Topic-guided Example Selection for Domain Adaptation in LLM-based Machine Translation

Seth Aycock, Rachel Bawden

Topic-guided Example Selection for Domain Adaptation in LLM-based Machine Translation

Seth Aycock[1,2*] **Rachel Bawden**[3]

[1]Institute for Logic, Language and Computation, University of Amsterdam
[2] Language Technology Lab, University of Amsterdam
[3] Inria, Paris, France
s.aycock@uva.nl rachel.bawden@inria.fr

Abstract

Current machine translation (MT) systems perform well in the domains on which they were trained, but adaptation to unseen domains remains a challenge. Rather than fine-tuning on domain data or modifying the architecture for training, an alternative approach exploits large language models (LLMs), which are performant across NLP tasks especially when presented with in-context examples. We focus on adapting a pre-trained LLM to a domain at inference through in-context example selection. For MT, examples are usually randomly selected from a development set. Some more recent methods though select using the more intuitive basis of test source similarity. We employ topic models to select examples based on abstract semantic relationships below the level of a domain. We test the relevance of these statistical models and use them to select informative examples even for out-of-domain inputs, experimenting on 7 diverse domains and 11 language pairs of differing resourcedness. Our method outperforms baselines on challenging multilingual out-of-domain tests, though it does not match performance with strong baselines for the in-language setting. We find that adding few-shot examples and related keywords consistently improves translation quality, that example diversity must be balanced with source similarity, and that our pipeline is overly restrictive for example selection when a targeted development set is available.[1]

1 Introduction

Adaptation of neural Machine Translation (MT) models to unseen domains remains a difficult problem because it requires handling out-of-distribution data at inference (Koehn and Knowles, 2017). Large language models (LLMs) offer an alternative method to the standard approach of fine-tuning an MT model or selected layers therein (Luong and Manning, 2015; Bapna and Firat, 2019). Openly available models such as Llama-2 (Touvron et al., 2023) and explicitly multilingual models including BLOOM (BigScience Workshop et al., 2023) and XGLM (Lin et al., 2022) perform well cross-lingually in classification and generation tasks, including many-to-many translation despite lacking explicit MT training.

However, regardless of the choice of LLM, some domains and vocabulary will remain under-exposed or unseen, especially for low-resource languages. Additionally, the optimal use of LLMs at inference to enhance translation quality remains under-explored. Domain adaptation of LLM-based translation is therefore an open and persistent challenge. Translation with LLMs requires prompting to elicit outputs in the desired language and domain, either via a zero-shot instruction or more effectively with in-context examples (Zhang et al., 2023a). In this work, we address the problem of domain adaptation at inference by exploring in-context example selection. Selecting lexically, semantically or grammatically relevant translation examples for prompting LLMs is arguably more important when translating out-of-domain texts, to help fill gaps in domain vocabulary or demonstrate different styles.

Many works select examples randomly from a development set (Brown et al., 2020; Chowdhery et al., 2022; Bawden and Yvon, 2023). However, other strategies have been developed. While some works show example diversity helps task performance (Zhang et al., 2022), intuitively we expect that examples showing translations of words in or related to the test source will improve output quality. Prior work has selected relevant examples based on n-gram overlap (Agrawal et al., 2023), feature matching (Kumar et al., 2023) or embedding similarity to the test source (Liu et al., 2022). Here we test a method that exploits more abstract semantic relationships that are also more

*This work was primarily carried out while at Inria.
[1]Our code, topic models, and data splits are available at www.github.com/Sethjsa/LLM-Dom-Ad.

fine-grained than domain categories. For this we use a topic modelling pipeline (Grootendorst, 2022) that predicts a topic for a source sentence, and selects examples from this topic for translating in the given domain. The motivation is two-fold: first, we aim to test the continuing relevance of these simple models to complement LLMs for MT; and second, we aim to study the importance of semantic similarity for domain adaptation at a more abstract granularity than prior work, providing an alternative method for example selection.

In practice we test Llama-2-13B, a state-of-the-art LLM, on MT in varied domains including medical, legal, educational, religious, and entertainment texts. We test across several high and low-resource languages from and into English: French, German, Czech, Romanian, Finnish, Lithuanian, and Tamil. We compare two uses of topic models for domain adaptation: topic-guided few-shot example selection and adding topic keywords. We test these against random baselines, information retrieval and embedding similarity-based selection, as well as simply adding domain labels. Our standard method uses multilingual topic models to select examples or keywords from seen domains across all tested languages. We show that our topic-guided method is robust to unseen domains and outperforms strong baselines in this setting, but is too restrictive to achieve competitive results against baselines for simpler in-language tests, suggesting a trade-off between similarity and diversity of examples.

2 Related Work

Domain adaptation methods for MT can be categorised as either data- or model-centric (Saunders, 2022). Data-centric approaches include fine-tuning models on in-domain parallel data (Dakwale and Monz, 2017) or synthetic backtranslated in-domain data (Sennrich et al., 2016; Jin et al., 2020), which is effective but costly in multilingual settings; or fine-tuning with labels encoding domain-specific information (Kobus et al., 2017; Stergiadis et al., 2021), which restricts prediction to seen domains. Model-centric approaches may use specialised architectures (Park et al., 2022) or different training methods such as curriculum or meta-learning (Zhang et al., 2019; Sharaf et al., 2020). Alternatively, adapters (Bapna and Firat, 2019) may be inserted into pre-trained models, with past work using separate domain and language adapters, or hierarchical domain adapters (Cooper Stickland et al., 2021a; Chronopoulou et al., 2022). Contrary to these approaches, we focus on domain adaptation of a pre-trained LLM at inference through in-context example selection, which requires no additional data manipulation or fine-tuning.

Recent work explores using pre-trained LLMs as a form of unsupervised transfer learning (Chronopoulou et al., 2020; Cooper Stickland et al., 2021b). Many LLMs are competent in multilingual translation despite lacking explicit MT training (Alves et al., 2023; Bawden and Yvon, 2023; Hendy et al., 2023; Peng et al., 2023), though LLMs often struggle in low-resource settings (Zhu et al., 2023). For zero-shot translation, prompt design is key, with prior work improving translation with instructions (Li et al., 2023), dictionary hints (Ghazvininejad et al., 2023), chained bilingual dictionary entries (Lu et al., 2023), or chain-of-thought prompting to predict keywords, topics, and relevant examples (He et al., 2023). In-context learning, i.e. providing few-shot task examples, is effective for LLM prompting (Brown et al., 2020), and various aspects of examples have been shown to impact translation quality: Vilar et al. (2023) find example quality outweighs domain provenance or source similarity, while Zhang et al. (2023a) show semantic similarity correlates with improved performance, and Zhang et al. (2022) show example diversity helps task performance more generally. Prior example selection methods include using n-gram-based BM25 retrieval plus a reranking model (Agrawal et al., 2023), a regression model with manually defined features to score retrieved prompts (Kumar et al., 2023), training a dense retrieval model (Rubin et al., 2022), or selecting based on proximity to the test source in a pre-trained LLM's embedding space (Liu et al., 2022). In this work we intend to achieve similar results using an alternative topic-guided selection method, permitting more abstract semantic relationships than n-gram overlap or embedding similarity.

Topic models are statistical tools that model latent semantic structure in texts (Blei et al., 2010), and while not state-of-the-art, these methods remain relevant for neural NLP. Prior work has integrated topic models into neural MT architectures to improve translation performance (Zhang et al., 2016; Wang et al., 2021), or fused external topic knowledge to improve domain robustness (Xezonaki et al., 2023). Aharoni and Goldberg (2020) study in-domain training data selection methods using unsupervised clustering methods based on pre-

`[Label]`	Domain: EU biomedical texts.
`[Keywords-10]`	Related keywords: stabilité, stabilumas, stability, lämpötilassa, temperatura, raumtemperatur, température, teplotě, temperatūrai, conservée.
`[Fewshot (1)]`	English: the lower operating value of ambient air temperature is minus 45 ° C; = Romanian: valoarea inferioară a temperaturii de funcționare a aerului ambiant – minus 45 °C;
`[Source]`	**English: Keep your Humalog Mix50 Pen in use at room temperature (below 30°C) for up to 28 days. = Romanian:**
`[Prediction]`	*Ștergeți penul Humalog Mix50 din uz la temperaturi de cameră (sub 30°C) pentru 28 de zile.*
`[Target]`	*Ţineţi Humalog Mix50 Pen în curs de utilizare la temperatura camerei (sub 30°C) timp de până la 28 zile.*

Table 1: An example illustrating our different prompting methods: domain labels, topic keywords, and a 1-shot example for an English–Romanian example from EMEA, with predicted and target outputs for the `Keywords-10` prompt. Information in square brackets is not included in the prompt.

trained language model embeddings, while Grootendorst (2022) introduces a neural topic modelling pipeline that clusters SBERT embeddings. We build on these works and train multilingual topic models to select relevant in-context examples for prompting at inference. Our work tests the continuing relevance of topic models in the context of example selection against information retrieval and embedding similarity baselines. The intuition is that topic models identify semantic relationships below the level of a domain but more abstract than embedding similarity or n-gram overlap. We expect this intermediate level of semantic abstraction will aid domain example selection for NMT.

3 Domain Adaptation Approach using Topic Modelling

3.1 Defining a domain

We employ topic models as the core mechanism of our domain adaptation approach. Traditionally in MT, a domain is defined as being a different source text, i.e. each corpus is taken as a different domain (Koehn and Knowles, 2017). Other definitions are more nuanced: Joshi et al. (2013) consider domains as consisting of multiple metadata attributes; van der Wees et al. (2015) subdivide domains into topic and genre characteristics; and Aharoni and Goldberg (2020) take a data-driven approach to defining domains, letting statistical models elucidate fine-grained cross-corpus associations and sub-domains within corpora. Building on the above, we suggest that domains can intuitively be defined by sets of distinctive words, forming a domain's vocabulary. We expect these words to be somewhat infrequent and pose a greater challenge for MT systems, suggesting this vocabulary should be prioritised for adaptation. Our definition, in addition to the data-driven approach, motivates using topic models for domain adaptation since they represent the latent semantic sets in a corpus.

3.2 Integrating domain information

Topic models find salient semantic relations between words or phrases in a corpus, representing these relations with a small number of abstract topics. Although typically modelled via the probabilistic latent Dirichlet allocation (Blei et al., 2003), we employ a different method which uses neural text representations as the basis for topics (Grootendorst, 2022). In these models, sentences are converted to contextual embeddings which are clustered based on similarity, then topic representations are extracted from these clusters of sentences using TF-IDF. Concretely, these topics consist of a set of associated vocabulary, and a set of representative sentences from the training corpus containing this vocabulary. We train multilingual topic models over data from seen domains in all languages on test. Once trained, any given input sentence can be embedded and assigned to the closest topic in the model. We therefore have three sources of additional information for each source sentence which we integrate into the translation prompt: a corpus-based domain label (`Label`), a list of keywords from the closest topic (`Keywords`), and a topic-guided set of representative examples (`Fewshot`). We illustrate our methods for integrating multilingual domain information in Table 1, with further examples in Appendix B.

Domain labels We add a descriptive domain label for a source sentence based on the corpus it comes from (i.e. following the standard definition of a domain as a corpus), referred to in results as `Label`. We avoid using the corpus name as these are not uniformly informative and instead use a short description (as shown in Table 3). We expect

this to slightly improve translation performance for given domains by conditioning the model to adapt to an expected style and topic.

Related Keywords For a given source sentence, we predict the closest topic from our model and use the 10 related keywords from that topic, referred to as `Keywords-10`. We hypothesise that adding keywords will marginally improve performance by both conditioning the model's generation context on the current domain by introducing distinctive domain vocabulary, and by acting as a stochastic proxy for a multilingual lexicon, often providing translations in other languages given the multilingual nature of the topic model.

Fewshot Examples Finally, we select the top n representative examples from a topic for a given source to use as in-context examples in the prompt, known in testing as `Fewshot` (n). We describe the variations of fewshot examples that are tested in Section 4. We expect topic-guided fewshot examples to result in larger performance improvements for these domains by showing semantically relevant vocabulary, grammatical sentence-level translations, and examples of the expected target domain style and output format. We also expect more examples to improve performance by giving further explicit translations. The topic model's ability to select semantically similar examples within domains, as opposed to random or n-gram matched examples, may allow the LLM to observe translations of domain-distinctive vocabulary, improving translation quality.

4 Experiments

Data and preprocessing We select several diverse high and low-resource languages: Czech (cs), German (de), English (en), French (fr), Finnish (fi), Lithuanian (lt), Romanian (ro), and Tamil (ta), both into and out of English. These languages vary from group 3 to 5 in Joshi et al.'s (2020) taxonomy of language resourcedness. We test on 7 domains with data in most languages: medical European Medicines Agency texts (EMEA), transcribed TED Talks (Reimers and Gurevych, 2020), localisation files for KDE4 software, educational video transcripts from QCRI (Abdelali et al., 2014) (QED), Quran translations (Tanzil), EU legal texts (JRC), and transcripts of TV and films from OpenSubtitles[2] (Lison et al., 2018) (Subs). All data was obtained from OPUS (Tiedemann, 2012), and we release our data splits, topic models, and code to aid reproducibility and future research.[3]

Our preprocessing involves removing newlines and sentences over 175 tokens with Moses scripts (Koehn et al., 2007);[4] removing sentences with over 50% punctuation; correct language identification using FastText (Joulin et al., 2017);[5] and sentence-level deduplication. For each domain-language pair, our development and test sets consist of 5000 and 500 sentences respectively (N.B. Tanzil Tamil–English has only 4800 dev set sentences). Development sets are used to train topic models and as sources for example selection.

Models We used the HuggingFace (Wolf et al., 2020) implementation of Llama-2-13B,[6] with greedy decoding up to 256 tokens. This model is mainly English with substantial multilingual capabilities, and is a state-of-the-art open-source LLM. We note however that the training data is not published, so our experiments are potentially at risk of data contamination. Our results therefore can only be considered in the context of this specific model.

While Llama-2 is not as performant on translation tasks as significantly larger models such as GPT-3.5 (Hendy et al., 2023; Xu et al., 2023), we chose Llama-2 for our experiments because it outperforms similarly-sized, explicitly multilingual LLMs including XGLM and BLOOMZ models (Zhang et al., 2023b). Llama-2 is also more robust to translation prompt perturbations than BLOOM models (Chitale et al., 2024). Further, Llama-2's permissive licence and open-source weights are a significant benefit against API-only models, leading to substantial research interest such as Llama-2-based translation models including ALMA (Xu et al., 2023). Our translation research on Llama-2 is therefore robustly motivated.

We implement our topic modelling pipeline with BERTopic (Grootendorst, 2022).[7] Our topic models are trained on parallel development sets; we focus on multilingual seen domain models, and also test in-language all-domain models. The multilingual setting is challenging and tests generalisation: without full domain coverage, methods must compensate, perhaps

[2] www.opensubtitles.org

[3] www.github.com/Sethjsa/LLM-Dom-Ad

[4] www.github.com/moses-smt/mosesdecoder

[5] www.fasttext.cc/docs/en/language-identification.html

[6] www.huggingface.co/meta-llama/Llama-2-13b

[7] www.github.com/MaartenGr/BERTopic

Template	Prompt
`Base`	**L1**: `[source sentence]` = **L2**:
`Verbose`	Given the following source text in **L1**: `[source sentence]`, a good **L2** translation is:
`Label`	Domain: `[domain description]` \n **L1**: `[source sentence]` = **L2**:
`Keywords-10`	Related keywords: `[keyword list]` \n **L1**: `[source sentence]` = **L2**:
`Fewshot` (n)	**L1**: `[example source]` = **L2**: `[example target]` \n **L1**: `[source sentence]` = **L2**:

Table 2: Prompt templates for our experiments. In each prompt, both the source and target language are specified to aid in the zero-shot setting. The bold **L1** and **L2** are replaced with full language names e.g. English or Lithuanian, and `[source sentence]` is replaced by a given L1 sentence. The `Fewshot` prompt includes n example pairs.

Domain	Description
EMEA	EU biomedical texts
JRC	EU legislative texts
KDE4	Software localization files
OpenSubtitles	TV and movie subtitles
QED	Educational video transcripts
Tanzil	Religious Quran text
TED	Public speaking transcripts

Table 3: Domain Label descriptions, providing similar information across domains. Bold domains are treated as *seen* in experiments.

by using related examples from non-target languages. We expect the abstract cross-lingual semantic relationships identified by topic models to show robustness across domains. All models use embeddings from the 100-language `paraphrase-multilingual-MiniLM-L12-v2` model from SentenceTransformers (Reimers and Gurevych, 2019), with UMAP dimensionality reduction (McInnes et al., 2018) and HDBSCAN clustering (Malzer and Baum, 2020). Our standard topic model has 500 topics trained on multilingual seen domains. Further details and hyperparameters are available in Appendix C.

Prompt design Our baseline experiments use the XGLM translation prompt (Lin et al., 2022), denoted as `Base`, and we also test a more verbose prompt, the two best performing MT prompts from Bawden and Yvon (2023) for BLOOM. Table 2 shows the format for our baselines and `Label` (see Table 3), `Keywords-10`, and `Fewshot` settings. Fewshot examples are selected from development sets of *seen* domains in all languages, unless otherwise specified; we expect this setting to be conducive to cross-lingual transfer.

Baselines Our `Base` setup uses a zero-shot XGLM-style prompt format. We implement two simple example selection techniques: BM25, an unsupervised information retrieval (`Retrieval`) technique based on n-gram matching; and sentence-level embedding similarity (`Similarity`), finding the closest sentences by cosine distance using the same SentenceTransformers model. For each baseline, we select from seen-domain multilingual development sets, and we additionally test in-language data against in-language topic models. Further baselines and ablations are described in Section 5. Finally, to contextualise our experiments we include topline results for NLLB-200-1.3B (Costa-jussà et al., 2022), a specialised translation model.

Evaluation We measure COMET scores using the `wmt22-comet-da` model (Rei et al., 2022), and BLEU (Papineni et al., 2002) with `SacreBLEU`[8] (Post, 2018), included for interpretability despite poorer correlation with human judgments (Mathur et al., 2020). We build on the `lm-evaluation-harness`[9] (Gao et al., 2022) for evaluation.

Postprocessing During initial tests we found the model often repeated outputs or provided translations in other languages, as found by Bawden and Yvon (2023). We therefore used a regular expression to capture the first L2 translation, discarding output after `(.+?:)` or a newline. Translation quality for trimmed results dramatically improves, showing the extent of Llama-2's overgeneration issues, and all results presented are trimmed outputs. In Appendix D, we show changes in length and correct language output for raw and trimmed results, plus `Base-raw` COMET and BLEU scores in Appendix E, confirming that trimming helps disentangle translation quality from overgeneration.

[8] `Signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`

[9] `www.github.com/EleutherAI/lm-evaluation-harness`

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
NLLB-1.3B	en–de	82.5	86.5	78.9	79.5	80.6	76.9	84.3
	en–ro	86.0	90.5	80.6	83.7	82.9	79.7	86.8
	lt–en	82.4	86.4	75.7	79.8	80.4	–	83.5
	mean	84.9	88.7	79.1	81.7	81.3	77.7	85.3
Base	en–de	71.8	74.0	72.1	75.0	74.4	68.5	78.2
	en–ro	66.5	79.2	70.9	74.1	72.5	65.7	77.1
	lt–en	62.4	62.5	55.6	54.3	56.7	–	58.6
	mean	70.8	73.5	67.4	68.2	67.2	62.6	70.0
Label	en–de	76.2	80.6	73.8	75.9	77.1	70.4	79.8
	en–ro	73.9	84.2	71.2	74.9	75.2	67.2	79.6
	lt–en	65.5	65.0	63.4	54.2	58.1	–	59.7
	mean	73.9	76.9	70.6	69.4	69.2	**64.1**	71.7
Keywords-10 (Seen)	en–de	77.6	81.2	75.8	75.9	77.4	70.1	80.0
	en–ro	76.7	84.3	74.8	75.9	76.5	67.2	80.4
	lt–en	67.4	67.1	63.7	56.9	59.4	–	60.4
	mean	75.3	77.7	71.3	69.6	**69.7**	63.6	72.1
Fewshot (3, Seen)	en–de	79.1	82.5	77.6	75.6	77.2	69.8	80.7
	en–ro	80.1	85.9	77.8	75.9	76.6	68.2	80.7
	lt–en	70.2	70.7	66.2	58.2	58.3	–	60.1
	mean	**77.0**	**79.4**	**74.5**	**70.3**	69.6	63.8	**72.5**

Table 4: COMET results for main experiments including domain labels (Label), 10 topic-guided keywords (Keywords-10), topic-guided 3-shot (Fewshot), from multilingual seen domains (Seen), and for topline NLLB model. Prompts are zero-shot unless specified; best performing mean results in **bold**; and a reference for experiment names is found in Appendix A.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Base	en–de	71.8	74.0	72.1	75.0	74.4	68.5	78.2
	en–ro	66.5	79.2	70.9	74.1	72.5	65.7	77.1
	lt–en	62.4	62.5	55.6	54.3	56.7	–	58.6
	mean	70.8	73.5	67.4	68.2	67.2	62.6	70.0
Label	en–de	76.2	80.6	73.8	75.9	77.1	70.4	79.8
	en–ro	73.9	84.2	71.2	74.9	75.2	67.2	79.6
	lt–en	65.5	65.0	63.4	54.2	58.1	–	59.7
	mean	**73.9**	**76.9**	**70.6**	**69.4**	**69.2**	**64.1**	**71.7**
Random Label	en–de	75.6	80.4	73.4	75.7	76.9	70.2	79.8
	en–ro	72.1	82.9	71.6	74.4	75.5	66.5	80.1
	lt–en	64.0	62.7	60.7	54.6	57.0	–	58.7
	mean	73.3	76.4	70.0	69.1	68.9	63.4	**71.7**

Table 5: COMET scores for Label prompts against Base zero-shot and Random Label prompts.

5 Results

Main Experiments We start by comparing our three approaches for integrating domain information into prompts, domain labels (Label), topic-keywords (Keywords-10), and topic-guided 3-shot examples (Fewshot), against our zero-shot prompt (Base) and the topline NLLB model. We also tested a verbose prompt to validate previous claims (Bawden and Yvon, 2023) and provide a point of reference, but since they are not central to our method, we present these results in Appendix E.

Table 4 shows COMET scores for a selection of representative high and low-resource language pairs plus mean results over all pairs.[10]

We first note that the baseline zero-shot Llama-2 model shows substantially reduced performance compared to the NLLB model, especially on the lower-resource languages of Romanian and Lithuanian, which is to be expected from a specialist MT model explicitly trained on these languages. NLLB is thus a useful topline for our experiments. However, fewshot prompting helps Llama-2 begin to approach the scores achieved by the NLLB model.

Experiments with a domain label in the translation prompt show increases in COMET scores of up to 3 points over the baseline model, though the effect is smaller for OpenSubtitles, perhaps due to its heterogeneity, and greater for Tanzil, for the opposite reason. This suggests the model is able to use this minimal domain information to condition the output style and improve translation quality in highly restrictive domains such as Tanzil.

Results for prompting with 10 related keywords show average improvements of 2-4 COMET over the baseline, and up to 1 point over domain label tests, except for the unseen Tanzil domain. This suggests that the topic model-predicted keywords are useful for the model, providing lexical information beyond a domain description, and acting as a proxy for a bilingual lexicon. We would expect a handmade bilingual lexicon to improve results further (Waldendorf et al., 2022) but we note that quality lexicons are rare and thus keywords from topic models are a useful approximation.

Topic-guided 3-shot examples provide the largest performance boost of up to 6 COMET points, outperforming keywords on all domains. Gains are smaller again for OpenSubtitles; and the unseen Tanzil and QED domain results are marginally outperformed by Label and Keywords results respectively, though performance remains competitive. This shows the difficulty of selecting relevant out-of-domain examples for domains with more distinctive vocabulary or styles. Overall these results support our hypothesis that while domain labels and keywords provide useful domain and lexical information, especially in restrictive domains, few-shot examples help the model to better mimic the task and produce the desired output format.

Our main results show substantial improvements in translation quality using our example selection

[10] For full COMET and BLEU results over all 11 language pairs, see Appendix E. We note BLEU scores follow patterns in COMET results.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Base	en–de	71.8	74.0	72.1	75.0	74.4	68.5	78.2
	en–ro	66.5	79.2	70.9	74.1	72.5	65.7	77.1
	lt–en	62.4	62.5	55.6	54.3	56.7	–	58.6
	mean	70.8	73.5	67.4	68.2	67.2	62.6	70.0
Keywords-10 (Seen)	en–de	77.6	81.2	75.8	75.9	77.4	70.1	80.0
	en–ro	76.7	84.3	74.8	75.9	76.5	67.2	80.4
	lt–en	67.4	67.1	63.7	56.9	59.4	–	60.4
	mean	**75.3**	**77.7**	**71.3**	69.6	**69.7**	63.6	72.1
Keywords-30 (Seen)	en–de	77.5	81.3	76.3	75.8	77.7	70.2	80.1
	en–ro	76.0	84.4	72.4	75.4	76.7	66.8	80.2
	lt–en	67.8	67.4	64.0	57.5	59.7	–	60.1
	mean	**75.3**	**77.7**	71.1	69.5	**69.7**	63.4	72.1
Keywords-10 (Seen, Random Topic)	en–de	76.5	81.3	75.4	76.0	77.3	70.2	80.1
	en–ro	75.4	84.7	73.0	75.6	76.4	67.3	80.0
	lt–en	64.4	64.8	61.8	54.9	57.6	–	59.2
	mean	74.2	77.4	70.7	69.5	69.5	63.5	71.9
Random Keywords-10 (Seen)	en–de	76.9	81.2	75.6	76.2	77.2	70.4	80.2
	en–ro	75.4	84.3	73.7	75.9	76.4	67.4	80.1
	lt–en	66.3	65.3	61.9	56.1	58.8	–	60.7
	mean	74.4	77.4	71.0	**69.8**	69.6	**63.8**	**72.2**

Table 6: COMET scores for topic-guided Keywords-10 and Keywords-30, random topic Keywords-10, and Random Keywords-10 from multilingual seen domains.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Base	en–de	71.8	74.0	72.1	75.0	74.4	68.5	78.2
	en–ro	66.5	79.2	70.9	74.1	72.5	65.7	77.1
	lt–en	62.4	62.5	55.6	54.3	56.7	–	58.6
	mean	70.8	73.5	67.4	68.2	67.2	62.6	70.0
Fewshot (3, Seen)	en–de	79.1	82.5	77.6	75.6	77.2	69.8	80.7
	en–ro	80.1	85.9	77.8	75.9	76.6	68.2	80.7
	lt–en	70.2	70.7	66.2	58.2	58.3	–	60.1
	mean	**77.0**	**79.4**	**74.5**	**70.3**	**69.6**	**63.8**	**72.5**
Fewshot (3, Seen, Random Topic)	en–de	79.0	81.5	76.9	75.3	77.1	69.8	80.1
	en–ro	79.2	85.0	76.0	75.5	76.9	67.6	80.8
	lt–en	59.8	60.7	52.3	49.5	51.5	–	53.5
	mean	75.1	77.1	70.9	68.4	68.5	63.5	71.2
Random Fewshot (3, Seen)	en–de	77.0	80.5	75.2	75.3	76.7	69.8	79.4
	en–ro	78.1	84.9	75.7	76.2	76.9	68.1	80.4
	lt–en	70.9	71.6	69.7	69.1	68.1	–	72.3
	mean	73.9	75.6	71.5	69.7	68.7	61.1	71.3

Table 7: COMET scores for Fewshot examples predicted by the multilingual seen-domain topic model, from one random topic, and random examples from seen domains across languages.

method. We now ablate each method against various baselines to further understand the source of these improvements.

Domain Labels While adding domain labels improves translation quality, we now test with randomised labels from the set of 7 labels to understand the source of improvements. The results in Table 5 show that, while prompting with the true domain label leads to overall better quality outputs across languages and domains, the random domain label tests produce similar improvements over the baseline, trailing the true label results by approximately 0.5 COMET points for most domains. This suggests that the presence of any additional structured information conditions the model to focus on the translation task, whether or not that information is directly useful for the current sentence.

Related Keywords We test various ablations of the Keywords prompt in Table 6. The Keywords-30 prompt is constructed using 10 keywords each from the top 3 predicted topics. Here we see equivalent or marginally lower quality compared to the standard Keywords-10 setting (less than 1 COMET point difference), suggesting that most gains stem from the first few keywords. The random topic setting adds keywords from a randomly selected topic from seen domains, and results are consistently lower than topic keywords, with results from Lithuanian to English showing substantial degradation across domains (up to −3 COMET points). This indicates related keywords may provide more utility in low-resource settings. Finally, Random Keywords-10 selects individual words randomly from multilingual seen domains, i.e. the topic model's training set. This setting is competitive with and on some domains outperforms the topic keyword prompts; while topic keywords provide semantically relevant words, and random topic keywords provide irrelevant but semantically consistent keywords, this setting provides genuinely diverse keywords, which appears to help performance. This suggests there is a trade-off between semantic relevance (through topic modelling) and information diversity in the prompt. In sum, only marginal gains can be attributed to the topic-guided method, suggesting the choice of keywords has less of an effect than the presence of keywords themselves.

Fewshot Examples We test topic-guided fewshot examples against random baselines, all 3-shot: fewshot examples from one random topic (Fewshot (Random Topic)), and random fewshot examples from seen multilingual data (Random Fewshot). The results in Table 7 show that while Fewshot (Random Topic) and Random Fewshot improve on the Base setting, the best results by 1-4 COMET points are achieved by the topic-guided example selection. This suggests that although there are gains to be had from simply adding random examples, the semantic relevance of these examples can lead to further improvements in translation performance. We expect this is due to a combination of both in-

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Fewshot (3, Seen)	en–de	79.1	82.5	77.6	75.6	77.2	69.8	80.7
	en–ro	80.1	85.9	77.8	75.9	76.6	68.2	80.7
	lt–en	70.2	70.7	66.2	58.2	58.3	–	60.1
	mean	**77.0**	**79.4**	**74.5**	**70.3**	**69.6**	**63.8**	**72.5**
Retrieval (3, Seen)	en–de	76.4	79.8	74.6	73.8	76.0	68.9	78.5
	en–ro	76.7	83.2	74.4	74.5	74.8	66.3	79.3
	lt–en	70.8	71.7	69.5	69.6	68.1	–	72.3
	mean	73.2	74.9	70.6	68.8	67.9	60.5	70.6
Similarity (3, Seen)	en–de	76.3	80.6	76.2	74.8	75.9	68.8	79.0
	en–ro	77.5	84.7	76.4	75.2	75.9	67.3	80.4
	lt–en	70.9	71.6	69.7	69.4	68.1	–	72.4
	mean	73.9	76.0	71.9	69.5	68.5	61.2	71.1

Table 8: COMET scores for Fewshot topic-guided examples, Retrieval selected examples, and embedding similarity selected examples (Similarity), all from seen domains across languages.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Fewshot (1, Seen)	en–de	78.1	81.3	75.9	74.8	76.3	69.0	79.8
	en–ro	78.5	84.6	75.5	74.7	75.5	67.6	80.1
	lt–en	64.0	68.3	62.4	55.8	57.3	–	58.1
	mean	74.8	78.5	72.3	69.3	68.9	63.5	71.6
Fewshot (3, Seen)	en–de	79.1	82.5	77.6	75.6	77.2	69.8	80.7
	en–ro	80.1	85.9	77.8	75.9	76.6	68.2	80.7
	lt–en	70.2	70.7	66.2	58.2	58.3	–	60.1
	mean	77.0	79.4	74.5	70.3	69.6	63.8	72.5
Fewshot (5, Seen)	en–de	79.5	82.4	78.2	75.8	77.4	70.3	80.1
	en–ro	80.4	85.8	78.3	76.1	77.1	68.9	81.2
	lt–en	70.9	71.2	67.9	58.1	59.0	–	60.3
	mean	**77.3**	**79.6**	**74.9**	**70.5**	**70.0**	**64.1**	**72.6**

Table 9: COMET scores for increasing Fewshot examples, from 1-shot to 5-shot, using our standard multilingual seen-domain 500 topic model.

creased embedding similarity and n-gram overlap. The topic model selects a topic probabilistically, and although there may be noise within the topic—for example, the representative sentence pairs are not always in the same language or the correct target language—the semantic cohesiveness of these sentences outweighs the noise present in randomly selected examples. We can therefore attribute a small amount of quality improvements to the proposed topic-guided method.

We also test our topic-guided selection method against strong baselines inspired by prior work: a Retrieval method using BM25, and an embedding Similarity approach. For both we select examples from multilingual seen domains to control the data available for selection, since our standard topic model was tested in this challenging set-up.

The results in Table 8 show competitive performance for Retrieval and Similarity baselines against each other. However, our topic-guided fewshot method achieves the best results across domains by up to 3 COMET points. We also see slightly larger improvements for tests on unseen domains (QED, Tanzil and TED). Our method is more robust to all three unseen domains since it relies on an intermediate level of semantic relations, more complex than n-gram overlap, more abstract than raw embedding similarity, and finer-grained than domain-level selection. The Retrieval baseline especially suffers in the unseen domains, underperforming or matching the baseline zero-shot setting for QED and Tanzil, we expect because with lower or zero vocabulary overlap, n-gram matching fails where embeddings can exploit contextual information.

Number of Examples We present topic-guided fewshot results in Table 9 for 1, 3, and 5-shot settings. The results show gains of circa 1 COMET point from 1-shot to 3-shot, and even smaller gains of approximately 0.3 COMET in overall performance from 3 to 5-shot. This suggests that 3 examples are sufficient to provide substantial translation improvements over a zero-shot baseline, with diminishing returns for adding extra examples, corroborating results for BLOOM (Bawden and Yvon, 2023). Lithuanian–English results show low-resource languages, especially those not in the model's training data, may benefit more from additional examples; here we see continued improvements from 1 to 3 to 5-shot.

Going Further We also test various topic model sizes. While most experiments use the multilingual seen-domain 500-topic model (trained on the devsets of 4 domains totalling 140,000 parallel sentences), we also experiment with 200- and 1000-topic models. The results in Table 10 are mixed; some domains exhibit improved performance with larger models, but the improvements for the 1000-topic model are small or negligible over the 500-topic model. This is unexpected; a larger topic model implies more semantic variety and thus a wider choice sentences to select for a given test source. However, we observed that in the Fewshot (3, Seen, 500 topic) setting across languages and domains, a 'general' catch-all topic is selected for 3.1% of tests, and the top 5 topics make up 16% of selected topics, when a uniform distribution would result in each topic having a 0.2% selection rate. Therefore the overselection of certain topics, and consequent reduction of sentences available for selection, is likely to reduce performance. We also

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Fewshot (3, Seen, 200 topic)	en–de	79.4	81.9	76.3	75.7	76.6	69.9	79.8
	en–ro	79.8	85.7	76.4	76.0	76.8	68.6	80.7
	lt–en	71.0	65.7	64.6	54.9	56.5	–	56.9
	mean	77.2	78.3	72.9	69.6	69.4	63.9	71.7
Fewshot (3, Seen, 500 topic)	en–de	79.1	82.5	77.6	75.6	77.2	69.8	80.7
	en–ro	80.1	85.9	77.8	75.9	76.6	68.2	80.7
	lt–en	70.2	70.7	66.2	58.2	58.3	–	60.1
	mean	77.0	79.4	**74.5**	**70.3**	**69.6**	63.8	**72.5**
Fewshot (3, Seen, 1000 topic)	en–de	79.0	82.3	78.1	75.6	76.6	70.4	79.7
	en–ro	79.8	85.8	77.0	75.9	76.6	68.4	80.1
	lt–en	71.2	70.9	64.6	56.3	57.8	–	59.7
	mean	**77.4**	**79.5**	74.1	69.9	**69.6**	**63.9**	72.0

Table 10: COMET scores for Fewshot examples from 200, 500, and 1000-topic seen-domain multilingual models.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Fewshot (3, Seen)	en–de	79.1	82.5	77.6	75.6	77.2	69.8	80.7
	en–ro	80.1	85.9	77.8	75.9	76.6	68.2	80.7
	lt–en	70.2	70.7	66.2	58.2	58.3	–	60.1
	mean	77.0	79.4	74.5	70.3	69.6	63.8	72.5
Fewshot (3, Language)	en–de	79.7	82.7	78.7	76.3	77.6	71.4	80.9
	en–ro	81.3	86.9	79.1	77.2	78.2	71.1	81.9
	lt–en	71.8	71.6	67.4	61.0	62.9	–	65.6
	mean	77.9	80.2	75.4	71.0	71.3	65.2	73.4
Retrieval (3, Language)	en–de	81.5	84.0	79.8	75.8	76.9	73.4	80.3
	en–ro	83.7	87.7	80.4	77.2	78.1	75.2	82.3
	lt–en	77.0	75.3	74.9	61.8	65.8	–	67.9
	mean	**81.2**	**82.5**	**78.8**	72.4	**73.4**	**71.0**	**75.3**
Similarity (3, Language)	en–de	80.5	83.2	79.3	75.9	77.9	73.8	80.2
	en–ro	83.5	87.6	79.9	78.2	78.3	74.0	82.7
	lt–en	76.8	75.9	73.2	64.9	67.3	–	69.2
	mean	80.6	82.2	77.2	**72.6**	72.9	67.5	74.5

Table 11: COMET scores for language-specific fewshot settings against our standard multilingual seen-domain Fewshot setup.

note that the homogeneity of examples within topics is likely to degrade performance since example diversity helps for other tasks (Zhang et al., 2022). Therefore issues remain with the restrictiveness of this method, which we leave open to future work.

In-language example selection Finally, we test language-specific 500-topic models to test whether the above results hold in a more restrictive scenario assuming the availability of in-language devsets in all domains, which is not always possible. Note here there are no unseen domains, though we include the multilingual seen-domain model for comparison. The results in Table 11 show improvements of 1-2 COMET for the all-domain language-specific models against the seen-domain multilingual model. We also provide language-specific Retrieval and Similarity results, which show even greater improvements compared to in-language topic-guided fewshot examples. Both baselines outperform the in-language topic model by 2 to 5 COMET, approaching the topline NLLB results in Table 4. This suggests that in the more restrictive scenario where we have full domain coverage in the target language pair, Retrieval and Similarity methods are very strong baselines because there is a greater probability of similar vocabulary, semantics, and syntax, while the topic model's noise and highly homogenous examples may hinder performance. However, in the more challenging scenario of the main results with unseen multilingual domains, our topic-guided method is more robust to domain shift.

6 Conclusion

We investigate the use of topic models for translation prompt construction and in-context example selection to aid domain adaptation for LLM-based MT. We train a multilingual topic model which, in a challenging multilingual seen-domain setting, outperforms random and statistical baselines, showing the importance of semantically similar examples. Our method offers a lightweight, robust solution for when no parallel data is available for a new domain. However if suitable (in-domain and in-language) development data is available then information retrieval and embedding similarity-based methods are more performant, simpler solutions. In future work, we intend to assess the transferability of our method to LLM-based translation more generally by testing across various LLMs, including more explicitly multilingual models. With this work we show an example of how statistical models can complement the performance of Llama-2, an English-centric LLM, in translation tasks to and from English.

Limitations

We recognise our work has limitations including: 1) We experimented with only one pre-trained LLM, Llama-2-13B. Further investigation is required to understand how our results and prompts would vary across a) different model families and b) varying model scales. We note therefore that our results are not generalisable to other LLMs, pending further work. 2) While we consider a variety of high- and low-resource pairs, all our tests are into or out of English. Further work is required to test other pairs, including both high–high and low–low resource pairs. We note that this may be more

difficult due to the reduced availability of good quality in-domain parallel data. 3) Our conclusions must be understood with the caveat that we do not know our chosen model's training data, including datasets and language distributions, beyond the basic information provided by Touvron et al. (2023).

Acknowledgements

Both authors' contributions were funded by Rachel Bawden's Emergence project, DadaNMT, funded by Sorbonne Université. Rachel Bawden's participation was also funded by her chair position in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001. The authors are grateful for the feedback provided by the anonymous reviewers.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context Examples Selection for Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised Domain Clusters in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering Large Language Models for Machine Translation with Finetuning and In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, Scalable Adaptation for Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Rachel Bawden and François Yvon. 2023. Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-

Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. ArXiv:2211.05100 [cs].

David Blei, Lawrence Carin, and David Dunson. 2010. Probabilistic Topic Models. *IEEE Signal Processing Magazine*, 27(6):55–65. Conference Name: IEEE Signal Processing Magazine.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Pranjal A. Chitale, Jay Gala, Varun Gumma, Mitesh M. Khapra, and Raj Dabre. 2024. An Empirical Analysis of In-context Learning Abilities of LLMs for MT. ArXiv:2401.12097 [cs].

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov,

and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. ArXiv:2204.02311 [cs].

Alexandra Chronopoulou, Matthew Peters, and Jesse Dodge. 2022. Efficient Hierarchical Domain Adaptation for Pretrained Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1336–1351, Seattle, United States. Association for Computational Linguistics.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2020. Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Asa Cooper Stickland, Alexandre Berard, and Vassilina Nikoulina. 2021a. Multilingual Domain Adaptation for NMT: Decoupling Language and Domain Information with Adapters. In *Proceedings of the Sixth Conference on Machine Translation*, pages 578–598, Online. Association for Computational Linguistics.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2021b. Recipes for Adapting Pre-trained Monolingual and Multilingual Models to Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. ArXiv:2207.04672 [cs].

Praveen Dakwale and Christof Monz. 2017. Fine-Tuning for Neural Machine Translation with Limited Degradation across In- and Out-of-Domain Data. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 156–169, Nagoya Japan.

Leo Gao, Jonathan Tow, Stella Biderman, Charles Lovering, Jason Phang, Anish Thite, Niklas Muennighoff, Thomas Wang, Zdeněk Kasner, Khalid Almubarak, Jeffrey Hsu, Pawan Sasanka Ammanamanchi, Dirk Groeneveld, Eric Tang, Charles Foster, Andy Zou, Ben Wang, Jordan Clive, Kevin Wang, Nicholas Kross, and Fabrizio Milo. 2022. A framework for few-shot language model evaluation.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation. ArXiv:2302.07856 [cs].

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. ArXiv:2203.05794 [cs].

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring Human-Like Translation Strategy with Large Language Models. ArXiv:2305.04118 [cs].

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. ArXiv:2302.09210 [cs].

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. A Simple Baseline to Semi-Supervised Domain Adaptation for Machine Translation. ArXiv:2001.08140 [cs].

Mahesh Joshi, Mark Dredze, William W. Cohen, and Carolyn P. Rosé. 2013. What's in a Domain? Multi-Domain Learning for Multi-Attribute Data. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 685–690, Atlanta, Georgia. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain Control for Neural Machine Translation. In

Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Aswanth Kumar, Ratish Puduppully, Raj Dabre, and Anoop Kunchukuttan. 2023. CTQScorer: Combining Multiple Features for In-context Example Selection for Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7736–7752, Singapore. Association for Computational Linguistics.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2023. Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions. ArXiv:2305.15083 [cs].

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot Learning with Multilingual Generative Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-Dictionary Prompting Elicits Translation in Large Language Models. ArXiv:2305.06575 [cs].

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Claudia Malzer and Marcus Baum. 2020. A Hybrid Approach To Hierarchical Density-based Cluster Selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 223–228.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Cheonbok Park, Hantae Kim, Ioan Calapodescu, Hyun Chang Cho, and Vassilina Nikoulina. 2022. DaLC: Domain Adaptation Learning Curve Prediction for Neural Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1789–1807, Dublin, Ireland. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Danielle Saunders. 2022. Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey. *Journal of Artificial Intelligence Research*, 75:351–424.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Amr Sharaf, Hany Hassan, and Hal Daumé III. 2020. Meta-Learning for Few-Shot NMT Adaptation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 43–53, Online. Association for Computational Linguistics.

Emmanouil Stergiadis, Satendra Kumar, Fedor Kovalev, and Pavel Levin. 2021. Multi-Domain Adaptation in Neural Machine Translation Through Multidimensional Tagging. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 396–420, Virtual. Association for Machine Translation in the Americas.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Marlies van der Wees, Arianna Bisazza, Wouter Weerkamp, and Christof Monz. 2015. What's in a Domain? Analyzing Genre and Topic Differences in Statistical Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 560–566, Beijing, China. Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Jonas Waldendorf, Alexandra Birch, Barry Hadow, and Antonio Valerio Micele Barone. 2022. Improving Translation of Out Of Vocabulary Words using Bilingual Lexicon Induction in Low-Resource Machine Translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 144–156, Orlando, USA. Association for Machine Translation in the Americas.

Weixuan Wang, Wei Peng, Meng Zhang, and Qun Liu. 2021. Neural Machine Translation with Heterogeneous Topic Knowledge Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3197–3202,

Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Danai Xezonaki, Talaat Khalil, David Stap, and Brandon Denis. 2023. Improving Domain Robustness in Neural Machine Translation with Fused Topic Knowledge Embeddings. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 209–221, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. ArXiv:2309.11674 [cs].

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting Large Language Model for Machine Translation: A Case Study. ArXiv:2301.07069 [cs].

Jian Zhang, Liangyou Li, Andy Way, and Qun Liu. 2016. Topic-Informed Neural Machine Translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1807–1817, Osaka, Japan. The COLING 2016 Organizing Committee.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum Learning for Domain Adaptation in Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1903–1915, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic Chain of Thought Prompting in Large Language Models. In *Proceedings of the Eleventh International Conference on Learning Representations*, Kigali, Rwanda.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. ArXiv:2304.04675 [cs].

A Results legend

In Table 12 we provide a brief reference for the experimental naming used in results tables.

Label	Description
`Base`	Zero-shot XGLM-style prompt
`Label`	Descriptive domain label + `Base` prompt
`Keywords-10`	10 related keywords selected from topic + `Base` prompt
`Fewshot`	Example source-target pairs selected by topic model + `Base` prompt
(1, Seen)	1-shot; selected from multilingual seen domains
(3, Language)	3-shot; selected from all-domain in-language data
(200 topic)	Selected using a 200-topic model.
(Random Topic)	Examples/keywords selected from one random topic.

Table 12: Reference for experimental terminology.

B Further examples

In Table 13 we provide further examples of the prompt format, and predicted outputs, for `Label`, `Keywords-10` (Seen), and `Fewshot` (3, Seen) settings.

C Topic Model Hyperparameters

We use UMAP and HDBSCAN implementations from cuML.[11] Our embedding model is `paraphrase-multilingual-MiniLM-L12-v2` from SentencePiece. This language model is a MiniLM model (Wang et al., 2020) distilled from XLM-R (Conneau et al., 2020) and thus is expected to have some knowledge of the 100 lan-

[11] www.github.com/rapidsai/cuml

[Label]	Domain: TV and movie subtitles.
[Source]	**Lithuanian: Šis miestas, ir viskas jame... = English:**
[Prediction]	*This city, and all of it...*
[Target]	*This city, everyone in it...*
[Keywords-10]	Related keywords: juice, nápojů, grapefruitsaft, štáva, grapefruit, drinks, greipfrutų, vartoti, pomerančová, frucht.
[Source]	**English: fruits — and they will be held in honour, = German:**
[Prediction]	*früchte — und sie werden in Ehren gehalten werden.*
[Target]	*Früchte, und sie werden geehrt*
[Fewshot]	English: (c) With effect from 1 July 1972 the text of Article 4 (2) and (3) shall be replaced by the following: = French: c) Le texte de l'article 4 paragraphes 2 et 3 est remplacé par le texte suivant, avec effet au 1er juillet 1972: English: 9. Article 28 shall be replaced by the following: = French: 9) L'article 28, est remplacé par le texte suivant: English: (h) the text of Part L. PORTUGAL shall be replaced by the following: = French: h) Le texte de la partie L. PORTUGAL est remplacé par le texte suivant:
[Source]	**English: a) in paragraph 1 the following subparagraph shall be added: = French:**
[Prediction]	*a) au paragraphe 1, le sous-alinéa suivant est ajouté:*
[Target]	*a) au paragraphe 1, l'alinéa suivant est ajouté:*

Table 13: Three examples illustrating our different prompting methods: domain labels, topic keywords, and a 3-shot topic-guided example from seen domains. We show examples for Lithuanian–English, English–German, and English–French in OpenSubtitles, Tanzil, and JRC domains respectively, with predicted and target outputs below the example prompts.

guages used in training. After dimensionality reduction and clustering, the inputs are tokenized using CountVectorizer and weighted with cTF-IDF (Pedregosa et al., 2011). The standard BERTopic hyperparameters are as follows: 500 topics, 10 keywords/topic, with stopwords removed for the vectorisation step; no stopwords were available for Lithuanian, so we used the top 100 most frequent words from our multi-domain development set.

Other parameters follow the standard implementation of BERTopic: UMAP: number of components = 5, number of neighbours = 15, metric = cosine distance; HBDSCAN: minimum samples=10. Finally we use the KeyBERT-inspired implementation to select the best 10 keywords to represent a topic, which avoids the repetitive selection of function words and stopwords. We also note here that predicting the nearest topic for a given input does not significantly slow down the inference process, with a rate of 80-100 iterations per second.

D Length and Language ID results

We present raw and trimmed results in Tables 14 and 15 for length and correct language identification respectively, for a selection of settings (`Base`, `Verbose` and `Fewshot` (3, Seen)). These results illustrate how the trimming procedure vastly reduces the length and improves the correct language identification of the outputs; note especially the high sentence lengths and low correct language identification for `Base-raw` experiments.

E Full COMET and BLEU results

We present full COMET and BLEU results in Tables 16–23, which follow the same patterns presented in Section 5.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Base-raw	en–cs	91.8	85.7	76.0	20.7	41.0	73.4	34.2
	en–de	51.5	95.2	69.2	15.1	32.4	32.2	28.9
	en–fi	94.5	89.3	77.7	29.7	40.1	–	44.0
	en–fr	58.3	82.8	74.6	24.2	44.3	63.5	34.2
	en–lt	99.2	84.6	69.9	25.2	53.4	–	57.0
	en–ro	83.7	87.8	65.7	19.6	36.4	57.7	35.5
	en–ta	–	–	57.6	14.1	50.5	27.3	22.8
	cs–en	27.8	43.9	28.3	7.9	14.8	31.3	16.0
	fr–en	28.5	45.8	32.2	8.8	17.0	25.3	18.2
	lt–en	29.9	52.8	31.4	7.8	18.0	–	19.1
	ro–en	28.7	44.9	31.1	9.8	15.9	19.1	17.5
	mean	59.4	71.3	55.8	16.6	33.1	41.2	29.8
Base	en–cs	16.7	25.0	6.4	7.8	12.0	17.3	14.4
	en–de	15.6	28.1	11.6	7.7	13.8	16.5	15.8
	en–fi	17.3	19.5	6.5	7.2	11.8	–	14.3
	en–fr	18.3	29.7	13.6	8.4	14.6	16.7	17.2
	en–lt	16.6	30.1	6.9	8.6	18.0	–	22.1
	en–ro	17.9	28.4	7.2	8.9	14.3	18.0	16.1
	en–ta	–	–	27.9	11.5	27.9	24.3	19.9
	cs–en	19.2	31.5	9.9	7.6	13.5	24.3	15.9
	fr–en	17.3	31.4	15.5	7.9	14.6	17.1	16.8
	lt–en	22.4	40.6	14.4	6.8	16.3	–	17.1
	ro–en	19.0	31.5	9.7	8.0	14.1	16.8	16.0
	mean	18.0	29.6	11.8	8.2	15.5	18.9	16.9
Verbose-raw	en–cs	16.7	26.4	6.7	8.7	13.4	17.7	14.9
	en–de	15.9	28.1	11.3	8.2	14.2	17.6	16.8
	en–fi	13.8	19.0	5.8	7.4	11.2	–	12.6
	en–fr	19.1	30.0	14.9	9.2	16.6	18.4	18.9
	en–lt	19.7	33.5	6.7	11.1	25.0	–	27.4
	en–ro	18.1	27.7	7.3	8.1	14.6	17.5	16.7
	en–ta	–	–	9.4	14.0	24.4	22.4	21.1
	cs–en	26.4	40.2	11.7	11.5	19.3	30.4	22.4
	fr–en	17.3	29.7	15.1	7.9	14.3	17.7	16.5
	lt–en	20.6	35.3	8.2	7.2	16.9	–	18.4
	ro–en	18.5	30.3	7.8	8.2	14.8	17.7	16.4
	mean	18.6	30.0	9.5	9.2	16.8	19.9	18.4
Verbose	en–cs	16.6	26.4	6.6	8.7	13.4	17.7	14.9
	en–de	15.8	28.1	11.2	8.2	14.2	17.6	16.8
	en–fi	13.6	19.0	5.8	7.4	11.2	–	12.6
	en–fr	19.0	30.0	14.9	9.2	16.6	18.4	18.9
	en–lt	19.7	33.5	6.7	11.1	25.0	–	27.4
	en–ro	18.1	27.7	7.3	8.1	14.5	17.5	16.7
	en–ta	–	–	9.4	14.0	24.4	22.4	21.1
	cs–en	26.4	40.2	11.7	11.5	19.3	30.4	22.4
	fr–en	17.2	29.7	15.0	7.9	14.3	17.7	16.5
	lt–en	20.5	35.3	8.2	7.2	16.9	–	18.4
	ro–en	18.5	30.3	7.8	8.2	14.8	17.7	16.4
	mean	18.5	30.0	9.5	9.2	16.8	19.9	18.4
Fewshot-raw (3, Seen)	en–cs	17.5	26.0	7.0	7.0	12.4	16.6	14.1
	en–de	18.1	27.0	13.3	7.8	14.1	17.5	16.4
	en–fi	14.6	19.7	5.9	6.5	12.5	–	12.3
	en–fr	20.1	29.1	16.2	9.0	16.2	17.8	18.0
	en–lt	18.4	29.9	6.1	8.9	20.9	–	23.2
	en–ro	19.3	27.6	7.1	8.0	14.4	17.1	16.3
	en–ta	–	–	8.3	10.0	23.1	21.4	18.1
	cs–en	18.2	30.4	7.2	7.4	12.7	21.7	15.8
	fr–en	17.3	30.0	14.4	7.7	14.4	17.3	16.4
	lt–en	17.5	31.4	7.2	6.5	13.8	–	15.3
	ro–en	18.5	29.7	8.0	7.8	13.9	16.7	15.8
	mean	18.0	28.1	9.2	7.9	15.3	18.3	16.5
Fewshot (3, Seen)	en–cs	14.7	24.3	6.1	6.8	12.0	16.6	13.9
	en–de	15.1	26.2	11.2	7.7	13.6	16.6	16.3
	en–fi	12.8	17.8	5.2	6.5	11.0	–	12.1
	en–fr	18.1	28.9	14.9	8.6	15.6	17.3	17.9
	en–lt	17.1	29.4	5.8	8.8	20.7	–	23.1
	en–ro	17.4	27.3	6.8	8.0	14.0	17.1	16.1
	en–ta	–	–	8.0	9.9	22.9	21.4	18.1
	cs–en	18.1	30.4	7.1	7.4	12.7	21.7	15.8
	fr–en	16.5	29.8	14.3	7.7	14.4	17.2	16.4
	lt–en	17.2	31.0	7.1	6.5	13.8	–	15.3
	ro–en	17.9	29.5	7.3	7.8	13.9	16.7	15.8
	mean	16.5	27.5	8.5	7.8	15.0	18.1	16.4

Table 14: Average length measured in space-tokenized words for selected settings.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Base-raw	en–cs	30.0	62.0	23.0	76.8	70.8	66.2	79.2
	en–de	57.2	58.4	40.6	92.0	86.0	87.6	88.6
	en–fi	12.8	31.6	11.4	56.8	49.2	–	53.0
	en–fr	44.4	47.0	21.2	73.8	67.6	55.6	80.0
	en–lt	13.6	42.6	23.4	67.0	48.8	–	52.2
	en–ro	24.6	60.8	28.8	80.4	79.8	70.4	81.6
	en–ta	–	–	48.4	78.4	43.0	87.6	79.0
	cs–en	91.8	91.6	89.2	97.8	98.4	97.4	99.6
	fr–en	95.4	95.2	95.0	99.2	99.2	99.0	99.8
	lt–en	88.4	91.0	86.0	94.6	96.2	–	95.4
	ro–en	92.4	91.4	89.6	97.6	98.0	98.6	99.2
	mean	55.1	67.2	50.6	83.1	76.1	82.8	82.5
Base	en–cs	41.4	79.8	59.8	82.8	78.4	90.0	83.4
	en–de	67.2	76.0	64.8	94.8	89.2	97.0	93.0
	en–fi	22.2	61.0	36.2	65.4	57.6	–	58.0
	en–fr	62.8	81.4	64.4	80.4	79.6	94.0	88.4
	en–lt	19.0	59.4	43.2	72.8	56.2	–	57.8
	en–ro	41.0	82.8	56.4	86.0	85.6	96.6	85.2
	en–ta	–	–	51.0	79.4	43.4	87.4	79.8
	cs–en	96.6	97.8	94.0	98.0	99.2	99.4	99.6
	fr–en	96.8	98.4	96.2	99.2	99.2	99.6	99.8
	lt–en	91.2	94.8	89.4	94.6	96.4	–	96.0
	ro–en	96.6	98.6	93.4	97.4	98.6	99.6	99.6
	mean	63.5	83.0	68.1	86.4	80.3	95.5	85.5
Verbose-raw	en–cs	95.0	98.8	89.8	93.0	96.0	97.8	98.2
	en–de	96.8	98.2	94.0	98.0	97.4	98.6	98.8
	en–fi	95.6	96.6	92.4	95.2	96.8	–	98.4
	en–fr	95.4	98.6	91.6	93.0	95.2	99.0	97.4
	en–lt	91.8	92.6	88.0	90.2	94.0	–	95.2
	en–ro	96.6	99.0	87.2	93.0	97.0	99.0	98.4
	en–ta	–	–	85.6	92.0	95.6	93.6	94.6
	cs–en	97.8	97.4	92.0	97.8	99.4	99.2	99.8
	fr–en	97.0	98.6	93.8	98.8	99.2	100.0	99.4
	lt–en	92.8	95.8	84.8	95.6	96.8	–	96.2
	ro–en	97.0	98.8	88.4	97.6	99.2	100.0	99.6
	mean	95.6	97.4	89.8	94.9	97.0	98.4	97.8
Verbose	en–cs	94.8	98.6	89.6	93.0	96.0	97.8	98.2
	en–de	96.8	98.2	94.0	98.0	97.4	98.6	98.8
	en–fi	95.4	96.6	92.4	95.2	96.8	–	98.4
	en–fr	95.2	98.6	91.4	93.0	95.2	99.0	97.4
	en–lt	91.8	92.6	88.0	90.2	94.0	–	95.2
	en–ro	96.6	99.0	87.2	93.0	96.8	99.0	98.4
	en–ta	–	–	85.6	92.0	95.6	93.6	94.6
	cs–en	97.8	97.4	92.0	97.8	99.4	99.2	99.8
	fr–en	96.8	98.6	93.6	98.8	99.2	100.0	99.4
	lt–en	92.8	95.8	84.8	95.6	96.8	–	96.2
	ro–en	97.0	98.8	88.4	97.6	99.2	100.0	99.6
	mean	95.5	97.4	89.7	94.9	96.9	98.4	97.8
Fewshot-raw (3, Seen)	en–cs	95.4	97.6	91.4	91.8	94.2	98.0	97.6
	en–de	95.4	99.2	92.6	99.4	98.4	97.8	99.4
	en–fi	95.4	96.0	88.4	96.4	95.2	–	98.4
	en–fr	94.8	99.0	92.8	92.4	94.4	98.4	98.0
	en–lt	93.8	97.2	86.4	90.2	93.8	–	94.8
	en–ro	96.4	99.4	85.8	92.8	95.4	99.0	98.4
	en–ta	–	–	95.2	98.0	96.0	99.8	98.8
	cs–en	98.2	99.4	94.2	98.2	99.4	100.0	99.6
	fr–en	98.2	99.8	96.6	98.8	99.6	99.8	99.6
	lt–en	96.2	98.4	90.0	95.8	98.0	–	98.6
	ro–en	97.4	99.4	91.4	98.2	98.6	99.4	99.6
	mean	96.1	98.5	91.3	95.6	96.6	99.0	98.4
Fewshot (3, Seen)	en–cs	96.0	98.2	92.0	92.0	94.8	98.2	97.6
	en–de	97.4	99.8	93.8	99.6	99.0	99.8	99.6
	en–fi	97.2	98.0	89.0	96.4	96.6	–	98.4
	en–fr	96.4	99.6	93.4	94.2	96.4	99.6	98.0
	en–lt	95.0	97.6	86.8	90.2	94.2	–	94.8
	en–ro	98.2	99.8	86.2	92.8	95.4	99.2	98.4
	en–ta	–	–	95.8	98.0	95.4	99.8	98.8
	cs–en	98.4	99.4	94.4	98.2	99.4	100.0	99.6
	fr–en	98.4	99.8	96.4	98.8	99.6	99.8	99.6
	lt–en	96.4	98.6	90.0	95.8	98.0	–	98.6
	ro–en	97.6	99.6	91.2	98.2	98.6	99.4	99.6
	mean	97.1	99.0	91.7	95.8	97.0	99.5	98.5

Table 15: Average correct language identification (%) measured with FastText's language ID tool.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
NLLB-1.3B	en–cs	88	91.8	81.8	84.1	83.4	77.6	86.4
	en–de	82.5	86.5	78.9	79.5	80.6	76.9	84.3
	en–fi	87	90.9	80.9	84.2	84.9	–	87.7
	en–fr	83.7	88.6	78	79.1	80.2	78.2	84
	en–lt	84.7	90.6	78.8	81.4	81.3	–	85.6
	en–ro	86	90.5	80.6	83.7	82.9	79.7	86.8
	en–ta	–	–	76.9	80.1	73.2	85.2	83.9
	cs–en	86	86.1	79.9	81.6	81.5	74.3	84.2
	fr–en	84.6	87.5	79.4	80.7	82.8	75.5	85.9
	lt–en	82.4	86.4	75.7	79.8	80.4	–	83.5
	ro–en	84.5	87.6	79.2	84	83.2	73.9	86.4
	mean	84.9	88.7	79.1	81.7	81.3	77.7	85.3
Base	en–cs	65.7	76.6	71.6	70.8	69.6	63.2	72.1
	en–de	71.8	74.0	72.1	75.0	74.4	68.5	78.2
	en–fi	65.4	68.8	68.4	70.3	70.3	–	71.3
	en–fr	75.2	79.6	72.9	72.6	73.9	68.7	78.9
	en–lt	54.7	45.9	59.9	52.0	47.8	–	47.3
	en–ro	66.5	79.2	70.9	74.1	72.5	65.7	77.1
	en–ta	–	–	39.5	46.9	35.9	28.9	37.9
	cs–en	82.4	81.3	77.3	76.3	77.8	66.6	80.7
	fr–en	82.7	83.9	78.0	79.2	81.2	73.2	84.8
	lt–en	62.4	62.5	55.6	54.3	56.7	–	58.6
	ro–en	81.4	83.7	75.7	78.3	79.0	66.2	83.6
	mean	70.8	73.5	67.4	68.2	67.2	62.6	70.0
Base-raw	en–cs	49.8	52.4	43.6	62.3	60.3	47.0	65.2
	en–de	61.7	56.9	51.6	68.0	67.3	63.3	72.3
	en–fi	50.4	46.6	44.2	61.1	61.5	–	63.8
	en–fr	61.3	60.8	48.0	65.4	64.6	52.4	72.3
	en–lt	42.9	38.5	41.4	46.2	42.5	–	43.0
	en–ro	48.9	55.5	45.4	66.0	64.0	50.6	69.1
	en–ta	–	–	34.2	44.4	34.7	31.4	37.7
	cs–en	75.7	75.3	64.1	70.5	72.5	64.0	76.6
	fr–en	75.3	77.5	67.6	73.3	75.8	69.4	80.3
	lt–en	58.4	58.9	48.7	51.3	54.4	–	56.6
	ro–en	74.6	76.9	62.9	72.4	74.6	64.4	79.2
	mean	59.9	59.9	50.2	61.9	61.1	55.3	65.1
Verbose	en–cs	77.9	82.1	75.2	71.5	74.4	66.8	77.7
	en–de	77.3	79.2	73.5	75.5	76.6	70.6	80.0
	en–fi	77.8	79.6	73.5	75.6	78.2	–	80.9
	en–fr	78.4	83.0	73.5	73.7	76.2	70.9	80.7
	en–lt	48.7	44.4	60.7	52.0	46.2	–	44.9
	en–ro	78.4	83.9	72.7	75.7	76.7	69.8	81.9
	en–ta	–	–	48.9	46.1	36.1	32.2	38.7
	cs–en	72.5	73.6	72.4	72.0	71.7	62.6	73.1
	fr–en	80.4	83.2	75.6	78.0	79.2	73.2	82.9
	lt–en	65.9	67.1	60.1	59.5	61.4	–	63.7
	ro–en	80.4	82.8	73.9	78.9	79.7	67.2	83.7
	mean	73.8	75.9	69.1	69.0	68.8	64.2	71.7
Label	en–cs	74.1	81.0	72.0	72.3	70.9	65.6	75.1
	en–de	76.2	80.6	73.8	75.9	77.1	70.4	79.8
	en–fi	72.1	77.3	70.9	75.0	76.4	–	78.3
	en–fr	79.2	83.7	73.7	75.2	76.7	69.7	81.2
	en–lt	49.7	45.5	60.1	51.7	46.8	–	44.0
	en–ro	73.9	84.2	71.2	74.9	75.2	67.2	79.6
	en–ta	–	–	49.3	47.9	38.5	29.3	39.6
	cs–en	83.1	82.4	81.3	77.4	79.0	68.6	81.9
	fr–en	83.1	85.1	80.7	79.7	81.8	73.9	85.3
	lt–en	65.5	65.0	63.4	54.2	58.1	–	59.7
	ro–en	82.1	84.0	80.0	79.3	80.3	68.1	84.4
	mean	73.9	76.9	70.6	69.4	69.2	64.1	71.7
Label-R	en–cs	72.6	80.5	72.9	72.0	70.9	65.7	74.5
	en–de	75.6	80.4	73.4	75.7	76.9	70.2	79.8
	en–fi	72.0	77.6	71.4	74.2	75.7	–	79.2
	en–fr	78.8	83.3	73.9	74.8	76.7	69.4	81.0
	en–lt	49.6	45.8	59.4	51.9	46.3	–	45.4
	en–ro	72.1	82.9	71.6	74.4	75.5	66.5	80.1
	en–ta	–	–	47.9	47.4	38.4	30.1	39.3
	cs–en	82.9	82.5	80.1	77.2	78.8	66.0	81.6
	fr–en	83.0	84.6	79.8	79.4	81.8	73.0	85.2
	lt–en	64.0	62.7	60.7	54.6	57.0	–	58.7
	ro–en	82.0	83.6	78.8	78.8	80.2	66.2	84.0
	mean	73.3	76.4	70.0	69.1	68.9	63.4	71.7

Table 16: COMET scores for various zero-shot translation prompts, and for zero-shot NLLB tests.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Keywords-10 (Seen)	en–cs	78.6	83.3	74.8	73.0	74.8	65.8	76.9
	en–de	77.6	81.2	75.8	75.9	77.4	70.1	80.0
	en–fi	75.5	79.5	73.8	73.8	76.8	–	80.5
	en–fr	80.4	84.2	76.1	75.7	77.3	69.8	81.7
	en–lt	48.8	44.9	61.2	51.1	45.4	–	43.8
	en–ro	76.7	84.3	74.8	75.9	76.5	67.2	80.4
	en–ta	–	–	46.1	47.1	37.3	30.8	37.9
	cs–en	83.2	82.9	79.5	77.5	79.3	66.3	81.6
	fr–en	83.3	85.0	80.1	79.3	81.7	72.9	85.2
	lt–en	67.4	67.1	63.7	56.9	59.4	–	60.4
	ro–en	81.9	84.2	78.5	79.6	80.5	65.7	84.2
	mean	75.3	77.7	71.3	69.6	69.7	63.6	72.1
Keywords-30 (Seen)	en–cs	78.2	83.0	75.6	72.7	73.9	65.9	76.9
	en–de	77.5	81.3	76.3	75.8	77.7	70.2	80.1
	en–fi	75.9	79.2	73.0	74.3	77.0	–	80.7
	en–fr	80.2	84.6	75.7	75.7	77.6	69.8	81.4
	en–lt	48.4	44.6	60.1	50.4	44.9	–	43.5
	en–ro	77.1	84.9	74.7	75.7	77.2	68.1	81.4
	en–ta	–	–	45.5	46.0	36.8	30.3	38.1
	cs–en	83.1	82.9	80.0	77.5	79.2	64.9	81.9
	fr–en	83.1	85.0	79.4	79.3	81.6	72.9	85.0
	lt–en	67.8	67.4	64.0	57.5	59.7	–	60.1
	ro–en	82.0	84.5	77.8	79.7	80.6	65.4	84.4
	mean	75.3	77.7	71.1	69.5	69.7	63.4	72.1
Keywords-10 (Seen, Random Topic)	en–cs	76.6	83.5	74.0	73.3	73.9	65.4	76.0
	en–de	76.5	81.3	75.4	76.0	77.3	70.2	80.1
	en–fi	73.9	79.3	72.8	74.2	77.2	–	80.4
	en–fr	79.7	84.0	75.8	75.5	77.3	69.8	81.6
	en–lt	47.8	44.4	60.6	51.7	46.0	–	44.0
	en–ro	75.4	84.7	73.0	75.6	76.4	67.3	80.0
	en–ta	–	–	45.7	47.3	37.9	30.5	38.3
	cs–en	83.1	82.6	80.2	77.4	78.7	66.2	81.7
	fr–en	83.0	85.0	79.8	79.4	81.6	72.9	85.0
	lt–en	64.4	64.8	61.8	54.9	57.6	–	59.2
	ro–en	82.1	84.2	78.6	79.7	80.2	65.4	84.3
	mean	74.2	77.4	70.7	69.5	69.5	63.5	71.9
Random Keywords-10 (Seen)	en–cs	76.7	83.0	75.8	73.5	74.2	66.3	77.0
	en–de	76.9	81.2	75.6	76.2	77.2	70.4	80.2
	en–fi	73.4	79.3	74.0	75.2	77.2	–	80.9
	en–fr	80.1	84.3	75.4	75.4	77.4	69.9	81.7
	en–lt	46.9	44.6	60.8	51.5	45.3	–	43.9
	en–ro	75.4	84.3	73.7	75.9	76.4	67.4	80.1
	en–ta	–	–	46.1	47.5	37.9	30.3	37.8
	cs–en	83.0	82.6	79.8	77.5	79.2	66.6	81.9
	fr–en	83.3	85.1	79.3	79.3	81.7	73.3	85.2
	lt–en	66.3	65.3	61.9	56.1	58.8	–	60.7
	ro–en	82.3	84.1	78.5	79.6	80.5	66.4	84.5
	mean	74.4	77.4	71.0	69.8	69.6	63.8	72.2

Table 17: COMET scores for topic-guided and random keyword prompts.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Fewshot (1, Seen, 500 topics)	en–cs	79.0	84.1	77.7	71.6	73.6	65.0	76.7
	en–de	78.1	81.3	75.9	74.8	76.3	69.0	79.8
	en–fi	76.1	81.4	75.8	74.7	77.4	–	79.7
	en–fr	80.1	85.2	75.4	75.2	76.1	69.8	80.6
	en–lt	47.8	46.0	62.0	51.7	45.2	–	42.8
	en–ro	78.5	84.6	75.5	74.7	75.5	67.6	80.1
	en–ta	–	–	49.6	47.9	36.1	30.2	39.0
	cs–en	82.2	83.5	80.3	77.3	79.1	66.9	81.5
	fr–en	81.5	85.6	80.3	79.4	81.7	73.1	85.3
	lt–en	64.0	68.3	62.4	55.8	57.3	–	58.1
	ro–en	81.2	85.1	80.0	79.4	80.0	66.1	83.8
	mean	74.8	78.5	72.3	69.3	68.9	63.5	71.6
Fewshot (3, Seen, 500 topics)	en–cs	80.5	84.9	80.8	73.7	74.9	66.4	77.2
	en–de	79.1	82.5	77.6	75.6	77.2	69.8	80.7
	en–fi	79.3	82.5	78.3	76.0	78.5	–	81.5
	en–fr	81.2	85.4	77.9	75.9	77.3	70.2	81.3
	en–lt	49.4	46.5	64.6	52.2	45.3	–	44.8
	en–ro	80.1	85.9	77.8	75.9	76.6	68.2	80.7
	en–ta	–	–	51.8	48.7	36.5	31.3	39.4
	cs–en	83.9	84.3	82.2	77.9	79.2	66.7	81.8
	fr–en	83.9	86.2	81.5	79.8	82.0	72.8	85.3
	lt–en	70.2	70.7	66.2	58.2	58.3	–	60.1
	ro–en	82.6	85.4	80.9	79.3	80.1	65.3	84.3
	mean	77.0	79.4	74.5	70.3	69.6	63.8	72.5
Fewshot (5, Seen, 500 topics)	en–cs	80.7	85.1	81.1	74.0	75.6	66.9	78.1
	en–de	79.5	82.4	78.2	75.8	77.4	70.3	80.1
	en–fi	79.7	82.6	79.0	77.3	79.0	–	81.5
	en–fr	80.9	85.5	78.2	75.9	77.4	70.3	81.9
	en–lt	50.4	46.9	64.5	52.0	45.5	–	44.7
	en–ro	80.4	85.8	78.3	76.1	77.1	68.9	81.2
	en–ta	–	–	51.4	49.1	37.2	31.6	39.0
	cs–en	84.2	84.3	82.3	77.8	79.3	66.1	81.7
	fr–en	83.9	86.4	81.8	79.9	82.0	72.9	85.4
	lt–en	70.9	71.2	67.9	58.1	59.0	–	60.3
	ro–en	82.6	85.5	81.2	79.7	80.4	65.6	84.3
	mean	77.3	79.6	74.9	70.5	70.0	64.1	72.6
Fewshot (3, Seen, 200 topics)	en–cs	80.6	83.0	79.8	73.5	74.3	67.2	76.7
	en–de	79.4	81.9	76.3	75.7	76.6	69.9	79.8
	en–fi	79.4	82.4	77.3	76.0	79.0	–	80.8
	en–fr	80.6	85.2	76.5	74.8	77.6	70.2	80.8
	en–lt	50.9	46.9	63.8	52.5	45.4	–	44.2
	en–ro	79.8	85.7	76.4	76.0	76.8	68.6	80.7
	en–ta	–	–	51.2	48.4	37.2	30.7	38.8
	cs–en	83.9	82.7	77.9	77.1	79.1	65.8	81.4
	fr–en	83.5	85.3	81.3	78.9	81.4	73.1	85.0
	lt–en	71.0	65.7	64.6	54.9	56.5	–	56.9
	ro–en	82.9	84.6	76.4	78.3	79.4	65.4	83.9
	mean	77.2	78.3	72.9	69.6	69.4	63.9	71.7
Fewshot (3, Seen, 1000 topics)	en–cs	80.9	85.0	80.2	73.0	74.4	66.2	77.6
	en–de	79.0	82.3	78.1	75.6	76.6	70.4	79.7
	en–fi	79.7	82.3	78.3	75.9	78.8	–	80.8
	en–fr	80.8	85.4	77.5	75.4	77.4	70.1	81.0
	en–lt	51.3	46.9	64.6	51.7	45.6	–	43.6
	en–ro	79.8	85.8	77.0	75.9	76.6	68.4	80.1
	en–ta	–	–	51.6	48.5	36.7	30.9	38.4
	cs–en	84.1	84.2	81.9	77.6	79.5	66.4	81.6
	fr–en	83.6	86.3	81.3	79.6	81.7	72.9	85.2
	lt–en	71.2	70.9	64.6	56.3	57.8	–	59.7
	ro–en	83.3	85.7	80.3	79.4	80.0	65.6	83.9
	mean	77.4	79.5	74.1	69.9	69.6	63.9	72.0
Fewshot (3, Seen, Random Topic, 500 topics)	en–cs	80.2	83.9	78.6	73.1	74.5	65.9	77.2
	en–de	79.0	81.5	76.9	75.3	77.1	69.8	80.1
	en–fi	78.6	80.1	77.4	75.4	78.3	–	80.2
	en–fr	80.5	84.9	75.7	75.1	76.7	69.7	81.4
	en–lt	49.1	45.7	62.2	50.7	45.2	–	43.2
	en–ro	79.2	85.0	76.0	75.5	76.9	67.6	80.8
	en–ta	–	–	49.7	47.6	36.2	30.8	38.7
	cs–en	82.2	81.7	76.9	75.7	77.3	65.8	80.2
	fr–en	82.4	84.8	79.1	78.3	80.9	73.1	84.6
	lt–en	59.8	60.7	52.3	49.5	51.5	–	53.5
	ro–en	80.5	82.5	75.5	76.6	78.6	65.4	83.0
	mean	75.1	77.1	70.9	68.4	68.5	63.5	71.2

Table 18: COMET scores for various topic-guided few-shot example experiments.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Random Fewshot (3, Seen)	en–cs	78.3	83.1	77.8	73.1	74.3	65.3	76.0
	en–de	77.0	80.5	75.2	75.3	76.7	69.8	79.4
	en–fi	76.4	77.8	76.7	75.0	78.2	–	80.5
	en–fr	80.0	84.5	76.0	75.2	76.8	69.5	81.0
	en–lt	50.7	47.3	63.9	52.6	47.6	–	45.7
	en–ro	78.1	84.9	75.7	76.2	76.9	68.1	80.4
	en–ta	–	–	51.5	48.7	39.5	31.2	39.7
	cs–en	76.4	73.7	74.3	74.0	71.8	57.8	75.2
	fr–en	76.2	77.1	71.7	73.0	73.2	65.7	77.1
	lt–en	70.9	71.6	69.7	69.1	68.1	–	72.3
	ro–en	74.6	75.0	73.6	74.2	73.0	61.4	76.7
	mean	73.9	75.6	71.5	69.7	68.7	61.1	71.3
Fewshot (3, Language)	en–cs	81.9	86.2	81.5	74.1	76.6	66.3	78.7
	en–de	79.7	82.7	78.7	76.3	77.6	71.4	80.9
	en–fi	80.4	82.6	79.4	77.1	79.6	–	81.8
	en–fr	81.3	86.2	78.7	75.8	77.5	71.5	81.6
	en–lt	51.5	49.6	67.3	52.8	47.4	–	45.0
	en–ro	81.3	86.9	79.1	77.2	78.2	71.1	81.9
	en–ta	–	–	53.2	48.9	42.3	33.7	39.7
	cs–en	84.4	84.3	81.2	78.1	79.9	67.0	82.3
	fr–en	83.6	86.6	81.7	79.5	81.7	73.7	85.4
	lt–en	71.8	71.6	67.4	61.0	62.9	–	65.6
	ro–en	83.2	85.8	81.2	80.2	81.1	67.3	84.5
	mean	77.9	80.2	75.4	71.0	71.3	65.2	73.4
Similarity (3, Language)	en–cs	83.9	87.2	83.2	74.6	78.0	71.0	79.2
	en–de	80.5	83.2	79.3	75.9	77.9	73.8	80.2
	en–fi	82.3	86.7	80.7	76.9	81.3	–	82.5
	en–fr	82.3	86.5	79.3	75.6	77.5	74.7	81.7
	en–lt	63.2	56.9	70.4	58.1	53.0	–	52.1
	en–ro	83.5	87.6	79.9	78.2	78.3	74.0	82.7
	en–ta	–	–	52.7	56.0	45.3	36.2	39.0
	cs–en	85.0	84.7	84.0	78.2	79.9	68.9	82.5
	fr–en	84.3	86.8	83.3	79.6	82.0	72.3	85.4
	lt–en	76.8	75.9	73.2	64.9	67.3	–	69.2
	ro–en	84.2	86.3	82.8	80.6	81.1	69.1	85.4
	mean	80.6	82.2	77.2	72.6	72.9	67.5	74.5
Similarity (3, Seen)	en–cs	76.2	82.0	77.8	71.9	73.6	64.7	74.6
	en–de	76.3	80.6	76.2	74.8	75.9	68.8	79.0
	en–fi	76.3	78.1	77.0	74.5	77.5	–	80.0
	en–fr	79.6	85.0	76.4	74.5	76.3	69.2	80.3
	en–lt	55.0	51.5	65.2	54.2	48.2	–	46.5
	en–ro	77.5	84.7	76.4	75.2	75.9	67.3	80.4
	en–ta	–	–	52.9	49.8	39.2	33.2	40.2
	cs–en	76.5	73.7	73.9	73.7	71.8	58.5	75.3
	fr–en	76.1	77.3	71.7	72.6	73.4	66.2	77.2
	lt–en	70.9	71.6	69.7	69.4	68.1	–	72.4
	ro–en	74.6	75.0	73.5	74.0	73.1	61.7	76.4
	mean	73.9	76.0	71.9	69.5	68.5	61.2	71.1
Retrieval (3, Language)	en–cs	85.1	87.3	83.2	74.9	77.6	72.7	79.2
	en–de	81.5	84.0	79.8	75.8	76.9	73.4	80.3
	en–fi	83.2	87.4	81.7	77.9	80.9	–	82.8
	en–fr	82.4	87.0	79.1	75.9	77.9	75.6	81.5
	en–lt	66.1	58.0	73.0	57.6	54.6	–	52.3
	en–ro	83.7	87.7	80.4	77.2	78.1	75.2	82.3
	en–ta	–	–	63.3	58.0	52.0	60.4	48.6
	cs–en	84.8	85.0	84.6	77.9	80.3	69.3	82.8
	fr–en	84.6	86.8	84.3	79.6	81.8	72.4	85.2
	lt–en	77.0	75.3	74.9	61.8	65.8	–	67.9
	ro–en	83.9	86.2	82.8	80.3	81.2	69.3	85.5
	mean	81.2	82.5	78.8	72.4	73.4	71.0	75.3
Retrieval (3, Seen)	en–cs	75.8	81.0	75.8	69.2	72.0	63.9	72.6
	en–de	76.4	79.8	74.6	73.8	76.0	68.9	78.5
	en–fi	75.0	76.9	75.0	74.0	75.7	–	79.0
	en–fr	78.9	84.0	74.5	74.4	76.3	68.6	80.1
	en–lt	51.3	47.1	62.7	51.8	46.9	–	46.0
	en–ro	76.7	83.2	74.4	74.5	74.8	66.3	79.3
	en–ta	–	–	51.5	48.8	39.1	31.6	40.2
	cs–en	76.3	73.2	73.9	73.7	71.7	57.8	75.2
	fr–en	76.0	76.8	71.5	72.4	72.9	65.7	76.9
	lt–en	70.8	71.7	69.5	69.6	68.1	–	72.3
	ro–en	74.6	74.9	73.4	74.2	72.9	61.3	76.5
	mean	73.2	74.9	70.6	68.8	67.9	60.5	70.6

Table 19: COMET scores for few-shot baseline experiments.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
NLLB-1.3B	en-cs	32.6	42.2	25.6	25.8	24.4	18.1	24.2
	en–de	34.6	39.8	24.9	25.8	28.3	30.7	30.1
	en–fi	27	24.8	22.9	19.4	22.8	–	21.9
	en–fr	40.4	52.1	36.8	30	37.9	30	38
	en–lt	25.3	37.9	16.8	17	22.5	–	21.3
	en–ro	41.3	35.5	26.8	30.7	27.4	23.8	30.7
	en–ta	–	–	12	16.5	4.7	9.7	6.9
	cs–en	46.7	50	31.8	35.2	31.6	25.4	34.1
	fr–en	46.6	56.1	38.3	33	38	24.3	38.7
	lt–en	36	50.2	25	30.1	32.4	–	33.3
	ro–en	49.4	54.4	33.9	41.9	38.4	22.7	42
	mean	38	44.3	26.8	27.8	28	23.1	29.2
Base	en–cs	8.6	15.4	17.0	8.9	10.0	4.0	11.5
	en–de	18.5	20.2	16.0	18.2	19.6	11.1	21.0
	en–fi	5.3	5.0	13.7	5.0	6.4	–	6.0
	en–fr	20.8	28.3	25.4	20.5	25.3	12.1	27.2
	en–lt	4.0	3.2	4.4	1.1	0.9	–	0.9
	en–ro	9.0	18.7	14.6	11.2	13.6	4.9	17.5
	en–ta	–	–	0.1	0.6	0.8	0.1	0.1
	cs–en	29.7	32.5	21.6	24.5	24.4	10.5	27.0
	fr–en	37.2	40.2	34.9	27.6	34.9	16.8	34.2
	lt–en	8.6	11.5	4.4	4.2	7.2	–	8.6
	ro–en	35.6	37.9	22.2	29.6	29.8	9.9	34.2
	mean	17.7	21.3	15.8	13.8	15.7	8.7	17.1
Base-raw	en–cs	1.5	4.4	1.2	3.0	2.7	0.9	4.6
	en–de	5.3	5.8	2.4	8.0	7.7	6.0	10.7
	en–fi	1.0	1.1	1.0	1.0	1.8	–	1.9
	en–fr	6.4	9.6	4.8	5.9	7.7	2.9	12.8
	en–lt	0.7	1.2	0.4	0.3	0.3	–	0.3
	en–ro	1.9	5.9	1.3	4.7	5.0	1.4	7.3
	en–ta	–	–	0.1	0.5	0.6	0.1	0.1
	cs–en	19.9	22.8	6.3	23.2	22.0	7.9	27.0
	fr–en	21.6	26.2	14.8	25.4	29.3	11.5	31.8
	lt–en	6.3	8.6	1.8	3.6	6.3	–	7.4
	ro–en	22.5	25.5	5.6	23.2	27.0	9.7	32.4
	mean	8.7	11.1	3.6	9.0	10.0	5.1	12.4
Verbose	en–cs	15.7	18.6	15.6	7.6	10.7	4.0	13.0
	en–de	22.2	24.0	18.2	14.8	19.8	11.6	20.9
	en–fi	11.2	7.4	14.0	5.2	10.4	–	10.8
	en–fr	28.1	33.3	32.8	17.7	25.6	11.3	27.3
	en–lt	3.9	4.2	5.0	0.8	1.1	–	1.2
	en–ro	20.7	21.2	17.0	12.2	15.3	4.5	19.1
	en–ta	–	–	0.6	0.5	0.4	0.0	0.2
	cs–en	20.6	25.6	14.5	13.6	16.6	7.3	19.1
	fr–en	33.1	37.6	31.6	22.9	31.2	15.2	31.2
	lt–en	10.7	13.6	7.4	4.9	8.3	–	9.0
	ro–en	33.4	34.9	21.7	24.9	27.3	10.7	32.6
	mean	20.0	22.0	16.2	11.4	15.2	8.1	16.8
Label	en–cs	13.8	19.4	20.5	12.1	11.2	4.4	13.7
	en–de	22.8	25.5	19.0	18.7	22.1	12.5	21.7
	en–fi	9.0	7.5	16.6	7.1	10.1	–	11.6
	en–fr	28.8	33.7	31.7	22.7	29.7	13.2	29.2
	en–lt	4.2	4.2	5.3	1.1	0.9	–	1.2
	en–ro	16.3	22.7	15.5	13.8	16.4	5.2	20.3
	en–ta	–	–	1.4	0.6	0.6	0.0	0.1
	cs–en	33.2	37.3	31.4	26.1	26.2	14.8	28.1
	fr–en	39.2	43.4	37.3	28.8	35.9	18.3	35.6
	lt–en	12.7	14.7	11.7	5.1	8.3	–	8.7
	ro–en	39.0	42.1	31.1	31.7	30.9	12.2	34.7
	mean	21.9	25.1	20.1	15.3	17.5	10.1	18.6
Label-R	en–cs	12.9	18.9	19.4	11.2	12.3	4.3	12.5
	en–de	22.1	25.5	18.1	18.1	21.1	12.3	22.4
	en–fi	9.1	7.9	17.3	6.5	10.0	–	11.1
	en–fr	28.0	33.3	30.8	21.5	28.9	12.8	30.1
	en–lt	4.3	4.2	5.2	0.7	0.9	–	1.4
	en–ro	15.1	21.5	16.2	13.7	16.6	5.0	19.5
	en–ta	–	–	0.7	0.6	0.5	0.0	0.1
	cs–en	33.1	35.3	29.1	25.3	25.4	9.6	27.2
	fr–en	38.5	41.0	37.8	28.1	35.6	16.3	34.8
	lt–en	11.2	13.5	11.2	4.8	7.6	–	8.5
	ro–en	38.3	39.0	27.5	30.8	30.6	9.5	34.9
	mean	21.3	24.0	19.4	14.7	17.2	8.7	18.4

Table 20: BLEU scores for various zero-shot translation prompts, and for zero-shot NLLB tests.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Keywords-10 (Seen)	en–cs	16.4	21.8	13.0	11.3	14.2	4.2	15.2
	en–de	23.7	26.4	18.7	18.4	21.6	12.3	22.4
	en–fi	10.6	8.0	13.4	6.3	11.6	–	12.6
	en–fr	31.1	34.8	33.9	22.8	29.4	13.0	30.8
	en–lt	3.4	4.3	5.4	1.0	0.9	–	1.3
	en–ro	20.9	23.5	21.0	14.5	17.3	5.1	20.2
	en–ta	–	–	0.5	0.8	0.5	0.1	0.1
	cs–en	32.7	35.9	22.6	26.2	26.3	10.1	27.2
	fr–en	38.8	42.0	35.1	28.2	36.1	16.5	35.1
	lt–en	11.3	16.1	8.9	5.7	8.6	–	9.1
	ro–en	38.9	41.1	27.9	32.6	31.0	10.2	35.7
	mean	22.8	25.4	18.2	15.3	18.0	8.9	19.1
Keywords-30 (Seen)	en–cs	16.5	21.2	13.7	12.0	13.2	4.3	15.8
	en–de	22.9	26.4	19.4	18.5	23.5	12.1	23.1
	en–fi	11.2	7.9	9.4	6.5	10.8	–	12.9
	en–fr	30.8	35.2	34.0	22.9	30.5	12.7	30.2
	en–lt	3.6	3.9	4.0	0.7	0.9	–	1.1
	en–ro	21.0	24.1	20.5	14.0	19.4	5.3	21.0
	en–ta	–	–	0.6	0.5	0.5	0.0	0.2
	cs–en	32.9	37.5	23.8	25.5	26.1	9.3	27.2
	fr–en	37.6	41.3	34.4	28.4	35.6	16.4	34.9
	lt–en	11.5	14.6	6.4	5.9	9.1	–	9.2
	ro–en	39.9	40.6	21.6	32.1	30.9	9.8	35.1
	mean	22.8	25.3	17.1	15.2	18.2	8.7	19.2
Keywords-10 (Seen, Random Topic)	en–cs	15.4	21.9	13.7	12.3	12.1	4.4	13.9
	en–de	22.1	26.3	18.4	18.6	22.3	12.0	22.1
	en–fi	9.8	8.3	14.5	6.5	11.5	–	13.1
	en–fr	29.2	34.7	33.3	24.0	30.5	12.5	31.5
	en–lt	3.9	4.2	5.7	1.2	1.1	–	1.3
	en–ro	19.5	23.3	17.9	14.8	17.6	5.1	19.8
	en–ta	–	–	0.5	0.7	0.5	0.0	0.1
	cs–en	32.8	35.7	22.8	25.9	26.1	9.9	27.7
	fr–en	38.6	41.9	33.9	28.3	35.8	16.2	35.3
	lt–en	10.7	14.0	6.4	4.9	8.4	–	8.6
	ro–en	38.5	40.0	28.0	32.2	30.9	9.2	34.8
	mean	22.0	25.0	17.7	15.4	17.9	8.7	18.9
Random Keywords-10 (Seen)	en–cs	15.5	21.7	19.1	11.6	13.8	4.6	14.0
	en–de	23.5	26.3	20.2	18.5	22.5	12.3	22.2
	en–fi	10.2	7.9	18.6	6.4	11.2	–	14.2
	en–fr	30.0	35.0	32.6	23.5	30.5	12.8	31.0
	en–lt	2.9	3.9	4.7	1.1	1.1	–	1.5
	en–ro	19.5	23.6	19.3	15.9	17.8	5.0	20.2
	en–ta	–	–	0.4	0.6	0.5	0.0	0.2
	cs–en	33.0	36.0	27.5	25.1	26.7	10.2	27.9
	fr–en	37.9	42.4	31.4	28.0	35.9	17.0	35.1
	lt–en	12.5	13.6	7.1	5.7	7.6	–	9.7
	ro–en	38.9	39.8	25.9	32.0	31.3	10.0	35.2
	mean	22.4	25.0	18.8	15.3	18.1	9.0	19.2

Table 21: BLEU scores for topic-guided and random keyword prompts.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Fewshot (1, Seen, 500 topics)	en–cs	16.9	21.5	23.0	11.0	13.6	4.3	13.6
	en–de	24.7	25.7	20.7	17.6	21.2	11.3	22.0
	en–fi	11.3	9.0	17.3	6.6	12.8	–	13.1
	en–fr	30.3	35.3	33.5	23.1	29.2	12.9	29.0
	en–lt	4.8	4.9	6.6	0.8	1.0	–	1.1
	en–ro	22.8	23.3	21.4	14.3	17.0	5.0	20.2
	en–ta	–	–	0.7	0.7	0.3	0.0	0.2
	cs–en	33.7	38.4	27.8	26.1	25.8	10.9	27.5
	fr–en	37.6	44.6	36.7	28.7	35.8	16.5	35.3
	lt–en	12.7	17.5	11.0	5.2	8.1	–	7.8
	ro–en	40.3	43.7	32.0	31.6	30.1	10.1	34.9
	mean	23.5	26.4	21.0	15.1	17.7	8.9	18.6
Fewshot (3, Seen, 500 topics)	en–cs	18.1	22.3	24.8	12.2	13.5	4.2	15.4
	en–de	25.9	27.0	21.7	19.1	23.4	11.9	24.0
	en–fi	13.2	9.3	20.9	7.5	13.4	–	13.4
	en–fr	32.0	36.7	35.2	23.8	29.2	12.6	30.8
	en–lt	5.2	4.9	7.8	1.2	1.3	–	1.7
	en–ro	25.0	26.4	23.5	15.4	17.1	5.5	21.4
	en–ta	–	–	1.2	1.0	0.5	0.0	0.3
	cs–en	36.5	39.6	32.5	27.4	25.5	10.6	28.0
	fr–en	41.7	46.4	39.7	29.7	35.7	16.7	35.4
	lt–en	16.5	19.2	13.1	6.7	8.4	–	9.2
	ro–en	41.0	45.3	32.9	31.9	30.8	9.6	35.6
	mean	25.5	27.7	23.0	16.0	18.1	8.9	19.6
Fewshot (5, Seen, 500 topics)	en–cs	20.4	23.2	23.6	11.9	14.5	4.4	15.3
	en–de	26.6	27.7	22.4	19.0	23.5	12.2	23.9
	en–fi	12.7	10.4	21.4	8.1	13.3	–	13.0
	en–fr	32.1	36.1	35.2	24.1	29.6	12.5	30.5
	en–lt	5.4	5.4	7.1	1.0	1.1	–	1.6
	en–ro	25.8	25.8	23.4	16.8	18.2	5.7	20.9
	en–ta	–	–	1.0	1.0	0.6	0.0	0.2
	cs–en	38.1	41.1	33.0	27.5	25.6	10.1	27.8
	fr–en	42.4	47.9	40.5	29.5	35.9	16.2	35.6
	lt–en	16.6	20.3	13.5	6.1	9.1	–	9.2
	ro–en	41.6	45.5	34.0	33.1	31.0	9.4	35.5
	mean	26.2	28.3	23.2	16.2	18.4	8.8	19.4
Fewshot (3, Seen, 200 topics)	en–cs	17.7	20.6	24.8	12.4	12.9	4.8	14.5
	en–de	25.8	26.0	22.4	18.3	22.9	11.7	22.7
	en–fi	12.4	9.8	19.6	6.4	13.7	–	12.4
	en–fr	31.7	36.0	34.4	22.9	30.1	12.8	30.5
	en–lt	6.3	6.1	6.9	1.3	1.1	–	1.3
	en–ro	24.5	25.2	22.7	16.4	18.8	5.6	21.8
	en–ta	–	–	1.3	0.6	0.5	0.0	0.1
	cs–en	38.4	38.0	28.2	25.8	25.6	9.6	27.6
	fr–en	41.8	45.1	40.1	28.8	35.0	17.1	34.9
	lt–en	17.3	16.0	11.5	4.8	7.8	–	8.3
	ro–en	41.4	44.1	25.8	30.6	30.8	9.2	34.4
	mean	25.7	26.7	21.6	15.3	18.1	8.8	19.0
Fewshot (3, Seen, 1000 topics)	en–cs	18.5	22.8	23.7	11.2	13.5	4.3	13.9
	en–de	26.0	27.1	21.3	17.4	23.0	12.0	22.8
	en–fi	12.9	10.9	20.9	6.9	13.1	–	12.9
	en–fr	31.5	35.7	35.0	23.4	30.5	12.8	30.7
	en–lt	5.4	5.3	6.7	1.0	1.2	–	1.2
	en–ro	25.0	25.5	23.5	16.9	18.3	5.4	20.3
	en–ta	–	–	1.1	1.0	0.4	0.0	0.2
	cs–en	38.2	40.4	32.4	26.3	25.9	9.7	27.5
	fr–en	41.4	46.8	40.0	29.2	35.2	15.6	35.1
	lt–en	17.6	21.1	11.9	5.0	8.1	–	8.8
	ro–en	43.0	45.6	32.5	32.4	31.6	9.7	35.2
	mean	25.9	28.1	22.6	15.5	18.3	8.7	19.0
Fewshot (3, Seen, Random Topic, 500 topics)	en–cs	17.8	21.6	23.6	12.6	13.2	4.5	14.7
	en–de	25.4	26.7	21.5	17.6	22.5	12.1	23.2
	en–fi	12.3	8.4	19.7	6.9	12.7	–	13.1
	en–fr	31.4	35.3	33.7	23.9	29.2	13.1	31.2
	en–lt	4.6	4.5	5.1	1.0	1.1	–	1.4
	en–ro	22.6	24.2	22.2	16.2	18.1	5.3	21.6
	en–ta	–	–	0.6	0.8	0.4	0.0	0.1
	cs–en	33.4	35.5	26.4	23.6	24.3	9.5	26.5
	fr–en	38.5	42.6	37.6	27.6	35.3	17.1	34.9
	lt–en	9.8	13.7	3.4	2.2	5.0	–	6.4
	ro–en	38.7	39.7	24.0	29.2	29.6	9.2	33.2
	mean	23.4	25.2	19.8	14.7	17.4	8.8	18.8

Table 22: BLEU scores for various topic-guided fewshot example experiments.

Prompt	Pair	Dataset						
		EMEA	JRC	KDE4	Subs	QED	Tanzil	TED
Random Fewshot (3, Seen)	en–cs	16.1	21.6	22.2	13.0	14.5	4.4	13.3
	en–de	23.0	25.8	19.3	18.1	21.6	11.8	23.3
	en–fi	11.6	7.6	19.3	7.4	12.4	–	12.6
	en–fr	30.3	34.6	33.1	23.6	28.4	12.2	29.9
	en–lt	4.9	4.8	7.8	1.1	1.3	–	1.8
	en–ro	20.9	24.0	20.9	16.0	18.7	5.1	20.8
	en–ta	–	–	1.4	1.0	0.8	0.0	0.2
	cs–en	5.3	6.3	8.1	3.4	1.9	0.2	1.4
	fr–en	6.3	6.6	7.1	3.7	2.8	0.5	2.7
	lt–en	4.4	3.5	5.1	0.8	1.2	–	0.9
	ro–en	5.8	5.5	7.0	2.3	2.6	0.1	2.8
	mean	12.9	14.0	13.8	8.2	9.7	4.3	10.0
Fewshot (3, Language)	en–cs	19.4	24.5	24.6	12.6	16.2	5.4	15.5
	en–de	26.9	27.4	23.0	18.4	23.4	13.1	23.8
	en–fi	13.3	10.2	23.0	7.2	14.8	–	14.8
	en–fr	34.1	38.0	36.3	24.2	30.3	15.3	30.1
	en–lt	6.0	7.6	8.5	1.5	1.5	–	1.8
	en–ro	28.5	29.1	24.1	18.6	19.7	7.6	21.7
	en–ta	–	–	1.2	0.8	1.2	0.4	0.4
	cs–en	38.2	40.3	31.0	26.5	27.8	14.0	28.7
	fr–en	41.6	47.5	39.8	29.1	35.9	18.4	35.6
	lt–en	17.5	22.5	14.5	7.3	10.9	–	12.8
	ro–en	42.0	45.8	34.1	32.4	32.0	12.8	36.7
	mean	26.8	29.3	23.6	16.2	19.4	10.9	20.2
Similarity (3, Language)	en–cs	29.6	28.9	27.8	13.0	16.1	19.0	16.6
	en–de	30.6	32.8	24.9	17.8	24.8	23.2	23.0
	en–fi	21.1	22.8	23.6	8.4	19.3	–	16.6
	en–fr	38.3	42.7	36.9	23.0	29.1	31.4	32.0
	en–lt	12.4	12.3	17.4	2.5	5.7	–	6.2
	en–ro	35.2	34.2	28.0	19.8	19.3	24.8	23.3
	en–ta	–	–	1.5	2.3	2.0	0.5	0.2
	cs–en	45.2	43.2	36.8	26.7	28.2	16.4	30.0
	fr–en	46.8	49.6	43.0	28.7	36.4	18.0	36.5
	lt–en	28.3	28.2	23.6	10.9	17.3	–	18.9
	ro–en	48.4	49.9	35.7	33.6	32.8	13.9	37.8
	mean	33.6	34.5	27.2	17.0	21.0	18.4	21.9
Fewshot (3, Seen)	en–cs	17.5	22.9	22.9	11.1	12.5	3.9	12.9
	en–de	24.3	27.0	20.4	17.4	20.9	11.5	21.8
	en–fi	12.5	12.1	21.3	6.3	11.9	–	12.7
	en–fr	30.1	37.0	33.3	22.0	28.5	11.9	29.3
	en–lt	6.9	7.5	12.4	1.9	1.3	–	1.4
	en–ro	24.0	25.6	22.0	15.7	17.0	4.9	20.1
	en–ta	–	–	1.8	1.1	0.8	0.0	0.1
	cs–en	5.4	6.1	7.8	3.0	1.8	1.6	1.5
	fr–en	6.4	6.9	8.0	2.8	3.8	1.3	4.3
	lt–en	4.4	3.8	5.1	0.7	1.2	–	0.9
	ro–en	5.8	5.1	6.6	2.2	3.0	1.0	2.3
	mean	13.7	15.4	14.7	7.7	9.3	4.5	9.8
Retrieval (3, Language)	en–cs	31.4	30.9	29.1	12.5	17.9	21.4	18.0
	en–de	34.0	35.3	26.4	18.8	23.2	23.1	23.5
	en–fi	24.0	25.5	27.0	9.2	19.3	–	18.3
	en–fr	39.3	43.9	39.7	23.0	31.4	34.3	30.8
	en–lt	15.8	14.0	18.8	2.7	6.5	–	7.4
	en–ro	36.7	35.6	30.1	19.9	19.5	28.9	22.4
	en–ta	–	–	6.6	3.9	4.5	25.1	6.2
	cs–en	44.7	44.7	38.7	26.6	28.9	16.9	30.0
	fr–en	47.3	50.3	45.2	29.3	36.0	17.7	36.8
	lt–en	27.9	28.5	27.4	9.8	16.8	–	18.9
	ro–en	48.3	49.9	40.6	33.0	32.8	14.3	38.4
	mean	34.9	35.9	30.0	17.2	21.5	22.7	22.8
Retrieval (3, Seen)	en–cs	15.0	20.4	19.3	9.7	11.6	4.3	10.9
	en–de	23.7	25.5	19.9	15.9	21.3	11.7	22.1
	en–fi	10.8	7.6	19.4	6.7	9.0	–	12.3
	en–fr	29.2	35.1	32.8	21.7	29.2	12.1	29.8
	en–lt	4.9	4.4	6.9	0.7	1.3	–	1.5
	en–ro	20.2	22.4	20.2	15.0	16.2	4.9	20.6
	en–ta	–	–	1.4	0.9	1.1	0.0	0.2
	cs–en	5.1	3.5	7.2	1.7	1.1	0.1	1.4
	fr–en	5.7	5.2	6.4	1.9	2.0	0.1	1.6
	lt–en	4.4	3.6	5.1	1.0	0.9	–	0.9
	ro–en	5.2	4.3	5.9	3.5	2.1	0.1	2.0
	mean	12.4	13.2	13.1	7.2	8.7	4.2	9.4

Table 23: BLEU scores for few-shot baseline experiments.